Accepted Manuscript

Dilution correction for dynamically influenced urinary analyte data

Johannes Hertel, Markus Rotter, Stefan Frenzel, Helena U. Zacharias, Jan Krumsiek, Birgit Rathkolb, Martin Hrabe de Angelis, Sylvia Rabstein, Dirk Pallapies, Thomas Brüning, Hans J. Grabe, Rui Wang-Sattler

PII: S0003-2670(18)30932-2

DOI: 10.1016/j.aca.2018.07.068

Reference: ACA 236169

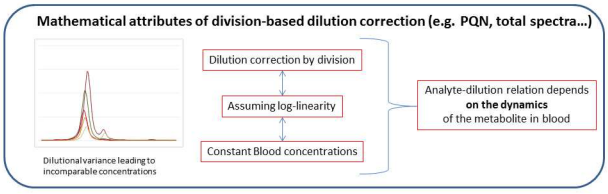
To appear in: *Analytica Chimica Acta*

Received Date: 28 February 2018

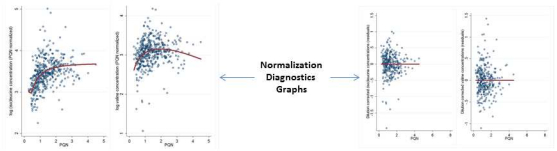
Revised Date: 29 June 2018

Accepted Date: 25 July 2018

Please cite this article as: J. Hertel, M. Rotter, S. Frenzel, H.U. Zacharias, J. Krumsiek, B. Rathkolb, M. Hrabe de Angelis, S. Rabstein, D. Pallapies, T. Brüning, H.J. Grabe, R. Wang-Sattler, Dilution correction for dynamically influenced urinary analyte data, *Analytica Chimica Acta* (2018), doi: 10.1016/j.aca.2018.07.068.



In **dynamically influenced data**, every analyte needs its **own dilution correction** dependent on its **time-course** in blood!



Division-based dilution correction
(biasing in dynamic data)

Regression-based dilution correction
(appropriate in dynamic data)

Dilution correction for dynamically influenced urinary analyte data

Johannes Hertel[1]*, Markus Rotter[2,3], Stefan Frenzel[1], Helena U. Zacharias[4], Jan Krumsiek[4,5], Birgit Rathkolb[5,6,7], Martin Hrabe de Angelis[5,8,9], Sylvia Rabstein[10], Dirk Pallapies[10], Thomas Brüning[10], Hans J. Grabe[1,11], Rui Wang-Sattler[2,3,5]

[1]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Germany
[2]Research unit of Molecular Epidemiology, Helmholtz Zentrum München, Germany
[3]Institute of Epidemiology, Helmholtz Zentrum München, Germany
[4]Institute of Computational Biology, Helmholtz Zentrum München, Germany
[5]German Center for Diabetes Research (DZD), München, Germany
[6]Chair for Molecular Animal Breeding and Biotechnology, Gene Center and Department of Veterinary Sciences, and Center for Innovative Medical Models (CiMM), Ludwig Maximilian University of Munich, Germany
[7]German Mouse Clinic (GMC), Institute of Experimental Genetics, Helmholtz Zentrum München, Germany
[8]Institute of Experimental Genetics, Helmholtz Zentrum München, Germany
[9]Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Germany
[10]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-Universität Bochum (IPA)
[11]German Center for Neurodegenerative Diseases (DZNE), Site Rostock/ Greifswald, Germany

Address for correspondence:
*Johannes Hertel
Department of Psychiatry, University Medicine Greifswald, Germany
Ellernholzstrasse 1-2, 17475 Greifswald
Tel.: ++49 – (0)3834/ 86 22 1 66; Fax.:++49 – (0)3834/86 68 89
e-mail: hertelj@uni-greifswald.de

Abstract

Urinary analyte data has to be corrected for the sample specific dilution as the dilution varies intra- and interpersonally dramatically, leading to non-comparable concentration measures. Most methods of dilution correction utilized nowadays like probabilistic quotient normalization or total spectra normalization result in a division of the raw data by a dilution correction factor. Here, however, we show that the implicit assumption behind the application of division, log-linearity between the urinary flow rate and the raw urinary concentration, does not hold for analytes which are not in steady state in blood. We explicate the physiological reason for this short-coming in mathematical terms and demonstrate the empirical consequences via simulations and on multiple time-point metabolomic data, showing the insufficiency of division-based normalization procedures to account for the complex non-linear analyte specific dependencies on the urinary flow rate. By reformulating normalization as a regression problem, we propose an analyte specific way to remove the dilution variance via a flexible non-linear regression methodology which then was shown to be more effective in comparison to division-based normalization procedures. In the progress, we developed several, easily applicable methods of normalization diagnostics to decide on the method of dilution correction in a given sample. On the way, we identified furthermore the time-span since last urination as an important variance factor in urinary metabolome data which is until now completely neglected. In conclusion, we present strong theoretical and empirical evidence that normalization has to be analyte specific in dynamically influenced data. Accordingly, we developed a normalization methodology for removing the dilution variance in urinary data respecting the single analyte kinetics.

Keywords: Normalization, dilution correction, urine analysis, metabolomics, model diagnostics, non-linear regression techniques

Abbreviations

iid= independent and identically distributed

ln= natural logarithm

OLS = ordinary least squares

PC ae = phosphatidylcholine acyl-alkyl

PQN = probabilistic quotient normalization

RCS = restricted cubic splines

SM= sphingomyelin

Vol_U = volume of urine produced between two urine voids

CV= coefficient of variation

LOD= limit of detection

1 Introduction

Urine, being easily accessible, is frequently used in life sciences, both in clinical and in basic sciences [1-3]. However, spot urine analyte concentration data is influenced by intra- and interpersonal variation due to the different dilutions [4-7], caused by variation in the velocity of urine production. The velocity of urine production, herein called the urinary flow rate, has a wide range of physiological values and changes strongly in response to challenges regarding the water balance of the organism (e.g. water or salt intake). As this variation among urinary concentrations is often not of interest, the data has to be normalized to remove the dilution variance according to the state of the art [4,8-11]. Different normalization procedures have been described in the literature, from osmolality normalization, specific gravity normalization and creatinine normalization to total spectra normalization or probabilistic quotient normalization (PQN) [9,10,12-15]. For most of these methods, a correction factor is derived [16] which is then supposed to be inversely proportional to the true urinary flow rate [17]. Hence, dividing the raw urinary analyte concentration by this factor should lead, in theory, to comparable urinary concentrations, relatively independent of the urinary flow rate. We call this approach within this paper *division based normalization*.

However, the implicit statistical and physiological assumptions behind the application of division were never systematically tested and analyzed. However, from the standpoint of statistical theory, the utilization of ratios of random variables to account for variability is known to be problematic [18]. Thus, this work aims at the clarification of the underlying assumptions behind division based normalization by transferring kidney physiology to a mathematical model. In a following empirical section, we deliver then a systematic test of our theoretical work, first on simulated data and then on real multiple time point metabolome measures.

In essence, we will show that division based normalization procedures are not adequate in dynamically influenced data-sets, because the application of division does not lead to the removal of dilution variance if the blood concentration and excretion kinetics are not time-invariant. Acknowledging the existence of analyte specific kinetics, we propose a different strategy to remove the dilution variance using flexible non-linear regression modeling. In the progress, we deliver several methods to check the validity of a certain normalization method which we call in analogy to regression diagnostics *normalization diagnostics*.

2 Theoretical Methods

As variation in the dilution is caused by the physiology of urine production, we will first explicate the renal processes influencing the relation between the urinary flow rate and concentration measures of freely filtrated analytes. The physiology is then expressed in mathematical terms such that we can explore analytically and in simulations the statistical dependencies induced by variation in the urinary flow rate. We will neglect for the simplicity of the main text tubular secretion. In Appendix A, however, we deliver a generalized model of the kidney including tubular secretion based on ordinary linear differential equations, leading to the same conclusions as the slightly less general model in the main text. We note that all mathematical modeling refers to the physiology of a healthy kidney which means in return that our model has to be treated with care when applied to pathophysiological states. In the following equations, time dependent variables are written in italic and bold small letters, while random variables and vectors of observations are displayed in bold and italic capitals.

2.1 Physiological relation between urinary flow rate and urinary concentrations under steady state assumptions

In a healthy individual around 90ml/min of blood are filtrated by glomerular filtration in the kidney [19]. Most of the fluid is then reabsorbed in different parts of the nephron such that

only around 1.5ml/min urine is actually generated in population average [20]. We call this velocity of urine production the urinary flow rate and denote it with $\boldsymbol{v}(\boldsymbol{t})$. This parameter, causal to the dilution variance [21], can vary strongly intra- and inter-individually. In the case of extreme fluid uptake, it can reach 11.2mL min^{-1} [22], whereas only urine output of less than 400mL per day (0.28mL min^{-1}) would be considered to be abnormal and an indication of kidney injury [23]. The urine is then stored in the bladder and gets released by urination if the stored volume approaches the capacity limits of the bladder, normally between 300mL and 400mL in an adult [24]. The higher therefore the urinary flow rate, the lower the time until the bladder reaches its capacity limits. This basic fact should be kept in mind in the following discussions.

Now, consider a freely filtrated analyte *y* with a time-invariant blood concentration $y_B = 100\mu mol\ L^{-1}$ such that 9µmol L^{-1} min^{-1} is filtrated. For many analytes, only a fraction of this filtrated amount is finally secreted due to reabsorption in the tubule. This fraction is called fractional excretion and is an analyte specific measure [25]. Consequentially, with a time-invariant urinary flow rate of 1.5mL min^{-1} and a fractional excretion of 5%, the urinary concentration produced per minute would be 300µmol L^{-1} min^{-1}. Under the strong assumption that every parameter is constant over time, the urinary concentration $\boldsymbol{y_U(t)}$ would be 300µmol L^{-1} at every time-point, because the excreted mass increases proportional with the excreted fluid.

In the next step, we formalize the exemplary calculations above, enabling a closer examination of the functional relations between the parameters. Technically spoken, we model the urinary concentration $\boldsymbol{y_U(t)}$ of an analyte *y* measured at the time $\boldsymbol{t}$ as the integral over the excreted, time-invariant blood concentration $\boldsymbol{y_B(t)} = y_B$ in the interval $[0, \boldsymbol{t}]$ between two urine voids:

$$(1)\ \boldsymbol{y_U(t)} = \frac{1}{\boldsymbol{vol_U}} \int_0^{\boldsymbol{t}} c_y \boldsymbol{y_B(s)} d\boldsymbol{s} = \frac{1}{v\boldsymbol{t}} \int_0^{\boldsymbol{t}} c_y y_B d\boldsymbol{s} = \frac{c_y y_B \boldsymbol{t}}{v\boldsymbol{t}} = \frac{c_y y_B}{v}.$$

Here, $\boldsymbol{vol_U}$ defines the volume of urine produced in the interval $[0, \boldsymbol{t}]$, c_y displays the excretion constant (product of filtration and fractional excretion) of the analyte under consideration, and *v* is the time-invariant urinary flow rate. It is important to realize that the length of the time-span $\boldsymbol{t}$ only gets cancelled out from the equation because we assume time-invariant parameters, leading to an inversely proportional relation between urinary flow rate and concentration. Note that inverse proportionality can be described with mathematical equivalence as log-linear relation, an attribute, we will use later on.

Now, if we normalize the urinary concentration by division, we assume that the utilized normalization factor (e.g. osmolality, PQN, creatinine, total spectrum) is inversely proportional to the urinary flow rate. Indeed, in the special case of (1) division will effectively remove the variance introduced by intra- and inter-individual differences in the urinary flow rate. However, this procedure is only valid under the strong assumptions of time-invariant parameters over the time-span between two urinations. As this assumption is likely violated in most natural settings, the question arises how the functional relations are shaped if we add dynamics to our modeling.

2.2 Physiological relation between urinary flow rate and urinary concentrations with dynamic blood concentrations

To study the consequences of non-constant blood concentrations for the functional relation between urinary flow rates and urinary concentrations, consider the simple example of a linear increase in blood concentration in the time interval between urinations. Thus, the blood concentration follows the function $\boldsymbol{y_B(t)} = a + b * \boldsymbol{t}$. Inserting the function in (1) gives

$$(2)\ \boldsymbol{y_U(t)} = \frac{1}{\boldsymbol{vol_U}} \int_0^{\boldsymbol{t}} c_y (a + b\boldsymbol{s}) d\boldsymbol{s} = \frac{c_y (a\boldsymbol{t} + 0.5 b \boldsymbol{t}^2)}{v\boldsymbol{t}} = \frac{c_y (a + 0.5 b \boldsymbol{t})}{v}.$$

We observe that in this case the time span *t* between the two urinations remains a factor in the equation. The time span however is a function of the urinary flow rate with an inversely proportional relation: $\boldsymbol{vol_U} = \mathrm{v}\boldsymbol{t} \leftrightarrow \boldsymbol{t} = \boldsymbol{vol_U}/\mathrm{v}$. Hence, we can rewrite (2) as

$$(3)\ \boldsymbol{y_U(t)} = \frac{c_y(a+0.5b\boldsymbol{t})}{v} = a\frac{c_y}{v} + b\frac{c_y 0.5\boldsymbol{vol_U}}{v^2}$$

The inverse proportionality we observed in the steady state model is no longer valid. In the consequence, normalization by division with a factor inversely related to the urinary flow rate will not remove the urinary flow rate from the equation assuming a linear increase in blood concentration.

Technically, as we integrate the blood concentration over a time interval depending on the urinary flow rate, the relation of the urinary concentration to the urinary flow rate is generally influenced by the analyte specific time course of its blood concentration. In Appendix A, the argument above is generalized to arbitrary blood concentration functions using an ordinary linear differential equation based model of the kidney, structurally similar to those that are used to model and control dialysis [26] and which contains the formulas for the calculation of fractional excretion [27] as a special case.

In the next step, we will discuss how we can remove the dilution variance from the data, respecting individual analyte excretion kinetics with arbitrary functional form.

2.3 Normalization as regression problem

To understand the solution we propose to deal with the analyte specific dependencies, it is important to notice that we can formalize dilution correction as a *regression problem* where we want to remove a variance fraction from the data. Thus, we enter the level of statistics and the parameters of the equation above have to be seen as *random variables* defined on a certain population. Common normalization strategies are based on

(4) $\boldsymbol{Y_{u_N}} = {}^{\boldsymbol{Y_U}}/_{\widehat{\boldsymbol{V}}} \Leftrightarrow ln\left(\boldsymbol{Y_{u_N}}\right) = \ln(\boldsymbol{Y_U}) - \ln(\widehat{\boldsymbol{V}})$

with $\boldsymbol{Y_{u_N}}$ being the normalized analyte concentration, $\boldsymbol{Y_U}$ the urinary concentrations, $\widehat{\boldsymbol{V}}$ an estimator inversely proportional to the urinary flow rate $\boldsymbol{V}$. Now, we define the regression problem (5) and its corresponding residual variable $\boldsymbol{Y_{res}}$ where we regress the log raw urinary concentration on the log normalization factor:

(5) $\ln(\boldsymbol{Y_U}) = \mathrm{b}_1 \ln\left(\widehat{\boldsymbol{V}}\right) + \mathrm{b}_0$ and $\boldsymbol{Y_{res}} := \ln(\boldsymbol{Y_U}) - (\mathrm{b}_1 \ln(\widehat{\boldsymbol{V}}) + \mathrm{b}_0)$

Constraining the regression coefficient with $\mathrm{b}_1 = 1$, the residual variable $\boldsymbol{Y_{res}}$ is equal to the log of the normalized urinary concentration plus a constant b_0. As adding constants does not change variances and covariances, normalization by division is statistically equivalent to deriving the residual of the displayed linear *constraint regression* [28] and $\boldsymbol{Y_{res}}$ and $\ln(\boldsymbol{Y_{u_N}})$ can be used interchangeable.

If the true relation between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{Y_U}$ is indeed log-linear with a slope of one, this procedure is adequate. However, as discussed above, the relation of the urinary flow rate to the urinary concentration is by no means restricted to log-linearity with slope one, but can display arbitrary analyte specific non-linearity. In this case, division based normalization (or linear constraint regression) will result in biased measures still dependent on the urinary flow rate. In essence, in dynamically influenced data, we have to estimate per analyte a potentially arbitrary non-linear dilution correction function. How to do so will be elaborated in the next section.

2.4 Regression based normalization using restricted cubic splines

As explained above, our goal is to derive an analyte specific dilution correction equation leading to the removal of the dilution variance in the analyte data. Arbitrary continuous

functions can be approximated via polynomials and polynomial equations can be estimated via ordinary least squares (OLS) procedures. Here, we suggest using restricted cubic splines (RCS) [29] to model flexible non-linear relations, but other methods like fractional polynomials may be used likewise. Restricted cubic splines, sometimes called natural splines, allow the estimation of piece-wise cubic functions, where the first and the last segments are restricted to be linear and the resulting function overall is forced to be smooth. The segments of the function are user-defined by specifying a set of *j* knot values k_1, k_2,…,k_j with $k_1<k_2\ldots<k_j$ and k_i within the range of the modeled variable. The corresponding OLS regression equation for j knots is given by (adapted from https://www.stata.com/manuals13/rmkspline.pdf):

(6) $\ln(\boldsymbol{Y_U}) = b_0 + \sum_{i=1}^{j-1} b_i f_i(\ln(\widehat{\boldsymbol{V}}))$ with $f_1(\ln(\widehat{\boldsymbol{V}})) = \ln(\widehat{\boldsymbol{V}})$ and for $i = 2,3,\ldots,j-1$:

$$f_i(\ln(\widehat{\boldsymbol{V}})) = \{(\ln(\widehat{\boldsymbol{V}}) - k_{i-1})_+^3 - (k_j - k_{j-1})^{-1}[(\ln(\widehat{\boldsymbol{V}}) - k_{j-1})_+^3(k_j - k_{i-1}) - (\ln(\widehat{\boldsymbol{V}}) - k_j)_+^3(k_{j-1} - k_{i-1})]\}(k_j - k_1)^{-2}$$

where $(u)_+ = u$ if $u > 0$ and $(u)_+ = 0$ if $u \leq 0$.

Restricted cubic splines are implemented in most statistical software, for example in R (https://www.R-project.org) in the inbuilt “splines”-package, so the user is not forced to apply (6) per hand (see for the R implementation the supplementary material).

As one can see from (6), restricted cubic splines have the advantage that the standard linear model is nested within the flexible non-linear model. Hence, if the true relation between dilution and analyte is log-linear with $b_1 = 1$ and $b_i = 0$ for $i = 2,3,\ldots,j-1$, the displayed regression model will deliver consistent estimates of this model. In this case, the corresponding residual variable will be a consistent estimator of the log-transformed division based normalized concentration. This means that division based normalization is a *special*

case of our more general regression-based framework. However, whereas division-based normalization implies time-invariance in blood concentrations and kidney parameters, the flexible non-linear regression based normalization does not rely on any special physiological state. It estimates an arbitrary functional relationship and is therefore not based on any specific model.

In conclusion, we propose that each analyte gets its own dilution correction which is estimated using flexible non-linear modeling in an OLS framework via the regression model (6). Normalization has then two steps:

1. Regress the log raw urinary concentration on the flexible non-linear modeled normalization factor (e.g. restricted cubic splines) for each analyte.
2. Derive the corresponding residual variable for each analyte.

The residual variable would then represent the dilution corrected urinary analyte measure which has by construction zero covariance with the utilized normalization factor. In contrast, in division based normalization, this attribute of zero covariance is not granted as we deal with a constraint regression. In drastic cases (e.g. strong deviations from steady state in blood), division based normalization may even increase the variance fraction explained by the urinary flow rate.

2.5 Statistical requirements for regression-based normalization

From a statistical standpoint, our methodology critically depends on independent and identically distributed (*iid)* observations as the whole sample is used to estimate the dependency of the urinary concentration on the urinary flow rate. This means that the observations are drawn independently from the same multivariate distribution of parameters. However, the methodology can be modified to deal with violations. For example, on repeated measurements, mixed models or general equation estimation procedures [30] could be used

for estimating the correction function, as observations from the same individual will show correlation between them. Another case of violation would be if a clinical phenotype changes the relation between urinary flow rate and urinary analyte concentration. In this case, it may be necessary to estimate the dilution correction function in a stratified way, allowing different estimates for different strata of the data (e.g cases and controls).

Note that we do not need any specific distributional assumption (e.g. log-normality) as the regression coefficients of OLS regressions are determined by the covariances and variances. It is however known that these can be largely influenced by outliers. In certain cases, it may be more appropriate therefore to use an outlier robust regression methodology, for example quantile regression [31].

2.6 Checking the efficacy of normalization: Normalization diagnostics

Until now, we did not explicate how to test whether an applied normalization technique (e.g. division-based or regression-based) effectively removed the dilution variance from the data. Different criteria were proposed in the literature [8,9, 12, 17], but they implicitly or explicitly rely on the log linearity assumption. Here, we propose to use graphical heuristics to choose the method of dilution correction, instead. These graphical assessments can be augmented with statistics if the graphical assessment does lead to ambiguous results.

2.5.1 Diagnostic graphs

Successful dilution correction should end in zero covariance of the analyte variables and the urinary flow rate. This can be tested graphically by plotting the dilution corrected analyte variable against the normalization factor. If any trend is perceivable, the dilution correction was not successful for this analyte. This is a very easy method already applied in a previous study [12] on metal ions, but is totally neglected in metabolomics. Likewise, the assumption of log-linearity can be easily tested graphically. One can plot the log of the raw analyte

concentration against the log dilution estimation and display additionally the identity (y=x) as reference line. If any analyte shows systematic departure from this line, division-based normalization will not work.

2.5.2 Diagnostic statistics

Sometimes, it may be hard to decide whether an observed trend is of statistical substance. In this case, we propose to test the constrained log-linear model (division based normalization) against the flexible non-linear model statistically. This can be done simply by regressing the division based normalized variables on the non-linear modeled (e.g. by restricted cubic splines) normalization factor. A significant model indicates then a systematic departure from log-linearity with a slope 1. This procedure here is asking whether the model-fit of the constrained log-linear model is equal to the model-fit of the flexible non-linear model. This comparison is here justified without further requirements because we test nested models against each other.

This test can be performed via parametric tests assuming normality of the residuals which is given if the metabolites are log-normally distributed. In this case, one can simply use the F-statistic of the corresponding regression. Alternatively, non-parametric tests could be performed using quantile regression or resampling methods.

2.6 Summary

To transfer our theoretical work to empirically testable hypotheses, we summarize the main points of the discussions above. First of all, our modeling is critically based on the assumption that normalization factors are inverse measures of the urinary flow rate and as such are related to the time-span between two urinations due to the limited volume of the bladder. This is the first testable hypothesis in a data-set where the time-span was sampled:

1. The time interval since the last urination is correlated with the normalization factors.

Secondly, in dynamically influenced data, we revealed that the time-span between urinations is a factor influencing urinary concentration measures, explicitly even after division-based normalization. Thus, we hypothesize:

2. The time-span between urinations is correlated with urinary analyte concentrations after division-based normalization.

Then, if the first two attributes are true in a data-set, it follows directly from our mathematical framework:

3. The normalization factors are not log-linear to the raw-concentration of urinary analytes and, hence, the urinary concentrations are associated with the normalization factors after division-based normalization.

The third hypothesis refers to the statistical attributes we use in normalization diagnostics. In combination, these three attributes lead to the central statement of this paper that in dynamically influenced urinary data, regression-based normalization is superior to division-based normalization in removing the dilution variance from the data.

3 Empirical methods

3.1 Simulations

The simulations were conducted to test first of all the PQN as estimator of the urinary flow rate in dynamically influenced urinary data, secondly, to illustrate the equivalence of regression-based normalization and division-based normalization in data derived under steady state assumptions, and thirdly, to test the ability of normalization methods to reconstruct the

true correlation matrix in data devoid of dilution variance. The simulations were designed according to the formulas in section 2.1 and 2.2 and the parameter settings were chosen to represent realistic measures of kidney physiology. A detailed description, a corresponding script in R, and an exemplary data-set can be found in the supplementary material. In essence, we vary the urinary flow rate in the context of a physiological system which is constrained by capacities and kinetic laws.

We derived seven data-sets with 500 observations simulating one steady state data set and six non-steady state data-sets with increasing dynamic range. For each of these seven data-sets, we additionally calculated a corresponding data-set with a fixed urinary flow rate of 2.5ml/min. This corresponding data-set devoid of variation in the urinary flow rate served as benchmark data-set. In essence, successful dilution correction should change the correlation matrix of the data such that it resembles the correlation matrix of the data-set devoid of variation in dilution.

Two normalizations based on the PQN factor were performed on each data-set with variation in the urinary flow rate: division-based and regression-based using restricted cubic splines. Additionally, we performed normalization by using the simulated urinary flow rate directly in division- and regression-based normalization. For the definition of the splines, four knots were used specified by the 5^{th} percentile, the 35^{th} percentile, the 65^{th} percentile and the 95^{th} percentile as recommended in [29]. Now, we calculated the distance (operationalized by the Frobenius norm, see supplementary material for details) of the correlation matrix of the differentially normalized data-sets to the correlation matrix of the benchmark data-set devoid of urinary flow rate variance. Then, the correlations of the urinary concentrations with each other resulting from the two normalization methods were calculated, along with the correlations of the normalized concentrations with the true urinary flow rate. Furthermore, exemplary cases of normalization diagnostic graphs were generated.

The whole procedure of the simulations was performed 500 times to assess random variation in the estimates of the Frobenius norm.

3.1 Study participants

For the practical application of the developed methodology and the test of the theoretical predictions, metabolome data from 100 female nurses between 25 and 65 years of age were used. The nurses were recruited at the clinical study site Bergmannsheil in Bochum, Germany. Exclusion criteria included 1) current pregnancy, 2) breastfeeding less than 6 months ago, 3) past or present fertility drug usage / medication, 4) prior cancer diagnosis. Each participant provided written informed consent. The study conformed to the Declaration of Helsinki and was approved by the Ethics Committee of the Faculty of Medicine of Ruhr-Universitaet Bochum (No. 4450-12).

From the 100 nurses, 75 nurses worked on night and day shift, while 25 nurses worked on day shift schedule only. During day shift, nurses worked for four consecutive days (Monday - Thursday) and up to five consecutive days (Monday - Friday) during night shift. Throughout day and night shift, urine samples and information on diet, sleep, and medication were collected.

3.2 Urine samples

Spontaneous urine samples were collected in 100 ml SARSTEDT disposable plastic containers, stored at 9°C for a maximum of 24 h before being aliquoted to 1.5 ml Eppendorf tubes and deep frozen at -80°C. The repeated measurement design led to urine samples collected over the whole 24 h range of the day, ensuring that dynamical influences should have an impact on the metabolome data. As each urine void led to one urine sample of which the time of day was noted, the time interval between urinations could be calculated.

3.3 Targeted metabolite quantification

Each urine sample was measured with the Absolute*IDQ*^{TM} p150 Kit (BIOCRATES Life Sciences AG, Innsbruck, Austria) using FIA-ESI-MS/MS (flow injection-electrospray ionisation-triple quadrupol mass spectrometry). The assay procedure using urine samples has been described previously [32]. In order to enable researchers to take into account different urine excretion rates, creatinine was included by the BIOCRATES Life Sciences AG into the metabolite panel.

In 10 µL urine, 162 metabolites were quantified for 2990 urine samples (Table S1). A urine pool based on a mixture of study participants' urine samples (positive control) was measured five times per 96-well plate, to calculate the coefficients of variance. In the course of quality control (QC), we excluded metabolites with a coefficient of variance (CV) higher than 25 %. Furthermore, three water based zero-samples per 96-well plate were measured to assess the limit of detection (LOD, which is defined as three times the median value of zero-samples). To ensure detectability we excluded metabolites with more than 50 % of measured values below LOD. Overall, 2990 urine samples were measured in two batches. The urine samples were randomly assigned to the batches, blocking samples from one individual with respect to the shift condition (night vs. day-shift). The QC was conducted separately for each metabolite for each batch. In total 44 metabolites passed the QC: free carnitine, 25 acylcarnitines (Cx:y), 13 proteinogenic amino acids, creatinine, hexoses (sum of hexoses), two phosphatidylcholine acyl-alkyl (PC ae), and one sphingolipid (SM) (Table S1). The abbreviations Cx:y depicts the total number of carbons and double bonds of all chains, respectively .

3.4 Osmolality

Freezing-point depression was used to determine osmolalities in "wakeup urine" samples which consisted of the morning urine samples at day shift and the first urine void after the

main sleeping period at night shift, resulting in 463 samples classified as wakeup urine samples. Osmolality measurements were made using a Gonotec Osmomat 030 (Berlin, Germany).

3.5 Normalizations

Several different normalized metabolome data-sets were derived. We utilized PQN, integral normalization, creatinine normalization, and osmolality normalization as described in the literature [12-14, 17]. For the PQN, the median concentration vector of all observations was used as reference spectrum. In sensitivity analyses, we used batch-specific PQN factors and person-specific PQN factors without seeing major differences in the results. Then, we modeled the normalization factors via restricted cubic splines using four knots at the 5^{th}, 35^{th}, 65^{th} , and 95^{th} percentile of the distribution of the corresponding normalization factor and regressed in mixed models the raw log concentrations of the metabolites on each of this non-linear modeled normalization factor. As the data contained repeated measurements per study participant, the regressions were designed as mixed models with random intercepts for the study participants. This implies the assumption of an exchangeable correlation structure which means that the correlations between measurements at different time points are supposed to be the same. Then, we calculated for each metabolite and each normalization factor the residual of these mixed effect regressions. Thus, in the end, we had eight (two for each method: division-based and regression-based) differently normalized data-sets for the wakeup urine samples, and 6 differently normalized data-sets in the whole sample, as osmolality measures were only available in the wakeup urine samples.

3.6 Statistical analyses

All metabolite concentrations were log-transformed. For sample description, urinary metabolite concentrations were summarized by mean values and standard deviations (see

supplementary Table S1). For statistical inference, mixed effect linear models were fitted using the study participants as random intercept variable, thereby respecting the repeated measurement design of the data. As mentioned before, this implies an exchangeable correlation structure over the repeated measurements. All other covariates were treated as fixed effect variables. Each analysis was run two-times, first for the whole sample including all available data and secondly for the subsample of wakeup urines. For statistical analyses of the time-span variable, we excluded all samples where the reported time-span between the two urine voids was longer than 12 hours for being potentially not reliable. In the supplementary material, we explicated all equations belonging to the described regression models above.

3.6.1 Testing the association of the time-span with the normalization factors

Here, we fitted a simple regression with the time span since last urination as response variable and the log of the normalization factor as predictor. This was repeated for each normalization factor.

3.6.2 Testing the association of the time-span with urinary analyte concentrations after division based normalization

Regarding the second hypothesis, we regressed each division-based normalized metabolite concentration (log-transformed) for each of the normalization factors on the time-span variable (non-linear via RCS using four knots as above), including the time of day (RCS with four knots), age, batch and body mass-index as covariates. The linearity of the time-span variable was tested by a Wald test, testing the coefficients of the second and third spline variable simultaneously on zero. As the first spline variable represents the linear trend, this procedure effectively provides a test on departure from linearity.

3.6.3 Testing the association of urinary concentrations with the normalization factors after division-based normalization.

The third hypothesis was tested analogously to the second hypothesis, with the time-span variable being exchanged for the log-transformed normalization factor (RCS with four splines) and the division-based normalized concentration being exchanged for the raw urinary concentration. The log-linearity was checked by testing the second and third spline variable derived from the normalization variable simultaneously on zero via a Wald test.

3.6.4 Normalization diagnostics

To deliver the diagnostic statistics described in section 2.5.2, we fitted the same models as explained above with the division-based normalized metabolite concentration as dependent variable and the corresponding normalization variable (RCS using four knots) as independent variable. Once again, exemplary normalization diagnostic graphs were generated.

4 Results

First, we describe the results of the simulation. Then, we move on to the results from the real metabolomic data.

4.1 Simulation results

The results of the simulation are comprised in Figure 1 and 2. The first relevant result is that even in dynamic data-sets the PQN factor was a fairly good estimator of the true urinary flow rate with a correlation of 0.99, regardless of the simulated dynamic range in blood (see Figure 1a and 1b). In the absence of dynamics, division-based normalization and regression-based normalization delivered statistically equivalent results and the residual variable correlated nearly perfectly (r=0.999) with the quotient of urinary concentration and PQN factor (see Figure 1c). However, with increasing dynamics both methods diverged from each other as

expected (see Figure 1d and supplementary Figure S1). This was mirrored in the differences between the "true" correlation matrix and the correlation matrices of the differentially normalized data-sets (see Figure 2A and 2B). While for division-based normalization the distance rises strongly with increasing dynamic range, there is only a very small increase in distance for the regression based methodology. Note that in Figure 2A the true correlation between the two simulated metabolites was zero; only using the true urinary flow rate in regression-based normalization was able to recover this trait from the data.

With larger dynamic range, we observed increasing inefficiency of division-based normalization to remove the urinary flow rate from the data (see Figure 2). After division-based normalization the urinary flow rate remained correlated with the normalized concentration (see supplementary Figure S2 and Figure 2D). For the regression-based approach, in contrast, the urinary flow rate was uncorrelated to the corresponding normalized concentrations regardless of the dynamic range in blood (see supplementary Figure S2 and Figure 2C). Thus, regression-based normalization provided in these simulations better results in comparison to division-based normalization in dynamically influenced data. Both methods are however statistically equivalent in steady state data-sets.

4.2 Results from urinary multiple time point metabolome data

Descriptive statistics of 2990 measured urine samples of 100 female nurses for 162 metabolites are shown in supplementary Table S1. Raw mean metabolite concentrations varied from 0.001 to 146.621 µM, displaying enormous intra- and inter-individual variance underlining the need of pre-processing steps for increasing the comparability of the data.

4.2.1 Empirical test of theoretical modeling

The first hypothesis of our theoretical work was that the time interval between urinations covariates with the normalization factors, regardless of the type of normalization factor. We

observed that all normalization factors were significantly associated with the time since last urination (log creatinine: regression coefficient b=1.10, 95%-CI:(1.00;1.20), p=3.05e-105; log PQN: b=1.31, 95%-CI:(1.18;1.44), p=1.83e-87; log integral: b=1.09,95%-CI:(0.98;1.19), p=9.03e-93) in the complete sample. This means that higher dilution was associated with shorter time-span between urine voids as expected. For visualization of the effects, see Figure 3. Even when restricting the analyses to the more standardized wakeup urine samples, the effect remained, regardless of the normalization factor utilized (log creatinine: b=1.53, 95%-CI:(1.17;1.89), p=1.33e-17; log PQN: b=1.83, 95%-CI:(1.38;2.27), p=5.83e-16); log osmolality: b=1.87, 95%-CI:(1.20;2.55), p=5.56e-08; log integral: b=1.56, 95%-CI:(1.19;1.92), p=1.00e-16).

For the second and third hypothesis, we only report the results for the PQN on the whole sample in detail in Table 2. Summaries regarding the other normalization factors (integral, creatinine, and osmolality normalization) can be found in the supplementary material. In general, all normalization factors behaved similarly, showing strong departure from the log-linearity assumption in most of the metabolites (see Table S2, supplementary material).

The second hypothesis was that the time-span is an important covariate for urinary data after division-based normalization. From 44 PQN (division-based) normalized metabolites surviving quality control, 38 were nominally significant. Bonferroni correction for multiple testing (corrected threshold: p=0.0011) would still lead to 25 significant metabolites (see Table 2) with 14 metabolites following a significant, metabolite-specific and non-linear trend as anticipated by the theoretical modeling.

The third hypothesis stated that the log raw concentration is not linear to the log of the normalization factor in dynamically influenced data-sets. From 44 tested metabolites, 40 showed a significant departure from log linearity regarding the PQN factor (34 after correction for multiple testing). Thus, for nearly all analyzed metabolites the normalization

factor was not log-linear to the concentration as could be expected from the results of the theoretical modeling.

Performing normalization diagnostic statistics, empirical data were in favor of the superiority of regression based normalization with 43 metabolites of 44 being related to the normalization factor after division-based normalization. This means that for nearly all metabolites division-based normalization did not remove the dilution variance completely. In the supplement, for each metabolite the gain in variance explained by non-linear modeling is listed in Table S3. Importantly, the dependency of the urinary metabolites on the normalization factor was metabolite-specific (see Figure 4 for an example from the wakeup urine data). Figure 3 is also an example for a normalization diagnostic graph.

We noted above that division-based normalization can *increase* the stochastic dependency on the dilution in drastic cases which would be contrary to our goal of removing the influence of the urinary flow rate. An example for this phenomenon can be seen in Figure 5. For PC ae 38:3, division-based normalization led to higher correlation (in absolute value) to the normalization factor compared to the raw urinary concentration. This was equally true for all other normalization methods based on division for this metabolite. For integral and creatinine normalization, we could observe the same for the acylcarnitines C16:2 and C18:2 (see supplementary Table S3).

4.2.2 The dilution variance as confounding variance

Here, we show that the urinary flow rate not only is a variance factor, but also an important confounding variable. In Table 2, we show for the example of PC ae 38:3 the association pattern to age and the sampling time using different normalization approaches. As creatinine and the integral normalization variable both were negatively associated with age (integral: b=

-0.014, 95%-CI: (-0.021;-0.006), p=0.0005; creatinine: b= -0.016, 95%-CI: (-0.023;-0.009), p=2.09e-05), the positive effect of age in the division-based normalized concentrations is a statistical artefact caused by residual confounding.

The anticipated higher power to detect real differences due to better removal of dilution variance can be seen in the results regarding the sampling time variable in intra-individual analyses. In the case of flexible non-linear regression-based normalization, the sampling time explained around 10% of intra-individual variance, while in division-based normalization the time variable explained not more than 3.08% of variance. This is a clear hint that division-based normalization was not as effective as regression-based normalization in removing intra-individual variance.

5 Discussion

In this work, we showed that the normalization of urinary data has to be seen in the light of a physiological system which is constrained by kinetic laws and limitations in capacities. Basically, from acknowledging that the bladder has a finite volume, it follows that dilution correction by division can only be efficient if all analytes under consideration have time-invariant concentrations in blood. Only in this very special case, the urinary flow rate is indeed inversely proportional to the urinary analyte concentration, as shown in our theoretical sections. This questions fundamentally the “state of the art” [4,5] of urinary data analyses as the metabolome is a highly dynamic system [33] influenced by nutrition [34], exercise [35], medication [36], the female cycle [37], and circadian rhythm [38]. Thus, the implicit statistical assumption (log-linearity between urinary flow rate and concentration with the same slope for all metabolites) behind the application of division cannot be expected to hold in data influenced by dynamic factors. This was mirrored by the provided results on simulated and real urinary data, showing clear analyte specific urinary flow rate dependencies. In

consequence, an analyte specific regression-based procedure allowing for arbitrary non-linearity resulted in a better removal of the dilution variance in simulations and in real data. Importantly, division-based normalization can be seen as a special case of the more general regression-based normalization we propose. Thus, as shown mathematically and in simulations, regression-based normalization will deliver equivalent results to division-based normalization if the assumptions behind division-based normalization are fulfilled.

We strongly believe that *normalization diagnostics* should be an important step in the analyses of urinary data, as it is a priori unclear which method of normalization is most efficient in removing the dilution variance from a sample. Generally, as the functional relation between urinary flow rate and concentration is dependent on dynamic factors, it is dependent in return on features of the study design. For example, it may be that for a highly standardized study the log-linearity assumption would be appropriate, but this has to be tested before applying a method based on log-linearity. Otherwise, the risk of wrong inferences and false positive results is not controllable, as demonstrated by the example of PC ae 38:3.

Our analyses showed additionally that the time-span since the last urination is an important covariate which is completely neglected in urine metabolomics until now. We think however that it is clearly of interest to assess the variation caused by, technically spoken, different integration windows over the blood concentrations. Considering non-constant blood analyte concentration, urine produced over a time-span of 2h will not be equivalent to urine produced over a time-span of 8h in its biological information. Thus, variation in the time interval variable may lead to non-generalizability of models, complicating the meaningful interpretation of results and hindering their transferal to clinical applications. In consequence, we suggest collecting data on the time-span because standardizing it will be impossible in most research settings, especially in large general population cohorts.

While our data is in alignment with our theoretical predictions from mathematical modeling, one has to recollect certain limitations of our work. A very important point to keep in mind is that the estimation of the potentially non-linear function is based on the *iid* assumption and does need certain sample sizes for a reliable estimation. In our case, we used restricted cubic splines with four knots which resulted in the estimation of three regression coefficients. On samples like the analyzed data-set comprising hundreds of urine samples, this is clearly no problem. In general, 10-20 observations per estimated parameter are considered to be sufficient for reliable estimates [29]. For small clinical samples therefore, it may be, in the sense of bias-variance trade-off, that division-based normalization procedures outperform regression-based normalizations while still introducing bias. Another limitation might be seen in the fact that we only analyzed metabolome data and, hence, our results may not generalize to other analytes like metal ions. However, our theoretical work explains also the pattern observed in [12, 39-41] describing the renal clearance of metal ions. Thus, we think our work is plausibly applicable to all kinds of urinary analytes which may not be at steady state in blood. Importantly, the mathematical model used here is likely too simple for real quantitative modeling as many aspects of human physiology were not respected. For example, for certain metabolites like hippurate, the secretion does not follow solely a first order kinetics due to tubular secretion [42]. Moreover, in individuals with diseases like diabetes or chronic kidney disease the model will be invalid, as for example metabolite concentrations may directly influence the urinary flow rate (e.g. glucose). Additionally, the model does not consider feedback mechanisms which may change statistical and functional relations between the parameters. Theoretically, on the other hand, regression-based normalization does not assume a certain physiological state. It estimates the relation between urinary flow rate and analytes from the data and is designed to cope with arbitrary functional relations, even in data reflecting pathophysiological states. This theoretical attribute of our methodology however has still to be tested in clinical data sets. Strictly spoken, the conclusions of this work are

therefore limited to healthy individuals. A final open question is how to perform the PQN on repeated measurement designs. In this study, we used the naïve approach of calculating the median of all samples, but it is not clear at all that this approach is best.

6 Conclusions

Normalization is a pivotal step in urinary data preprocessing which has been largely performed without clarifying the implicit statistical and physiological assumptions behind it. Here, we showed that division-based normalization is theoretically and empirically not valid in dynamically influenced data-sets. The conclusions of our work have important implications for studies and diagnostic procedures based on urine analysis, as the usually applied division-based normalization can introduce severe bias in dynamically influenced data-sets and therefore cannot be recommended. According to our work, regression-based normalization, as proposed above, will often be superior, enabling analyte specific dilution corrections. In conclusion, this work demonstrates the importance of understanding the statistical representation of dynamics in analyte concentration data and incorporating physiological constraints into the design of statistical analysis. Although challenging, we see serious potential in future work along these lines, enhancing the interpretability and the utilization of urinary concentration data.

Appendix A: The analyte specific functional dependency on the urinary flow rate in mathematical terms

Here, we will explicate the arguments for an analyte specific functional dependency on the urinary flow rate in mathematical terms. Consider the basic differential equation (A1) describing the change of a blood analyte level

(A1) $$\boldsymbol{y_B}'(\boldsymbol{t}) = \sum_{k=1}^{K} \boldsymbol{g_k}(\boldsymbol{t}) - [(1 - c_R)(c_F + c_S)]\ \boldsymbol{y_B}(\boldsymbol{t})$$

with $y_B(0) = y_0,\ \boldsymbol{y_B}(\boldsymbol{t}) > 0\ \forall\ \boldsymbol{t} \in \mathbb{R}$ and $c_R, c_F,$ and $c_S \in \mathbb{R}^+$,

where $\boldsymbol{y_B}(\boldsymbol{t})$ is the blood concentration of an analyte *y* as a function of time. $\boldsymbol{g_k}(\boldsymbol{t})$ denotes the sum of uptake and secretion into blood from all tissues k=1,…,K. Here, we understand the microbiota in the gut as an additional tissue contributing to the blood concentration. c_F describes the exponential removal of the analyte by filtration, c_S and c_R denote tubular secretion and reabsorption, respectively. By this formula, we assume first order kinetics for all transport processes and we define $c_y := [(1 - c_R)(c_F + c_S)]$. Note that equation (A1) is structurally similar to those equations that are used to model and control dialysis [26]. y_0 is the starting value at the time of last urination. The time of last urination was set to zero. The dynamics of the equation are largely driven by the tissue dependent component $\boldsymbol{g}(\boldsymbol{t}) = \sum_{k=1}^{K} \boldsymbol{g_k}(\boldsymbol{t})$. Here, we are only interested in understanding the effects of dynamics in the relation to the urinary flow rate. For other applications like quantitative modeling, the functions $\boldsymbol{g_k}(\boldsymbol{t})$ would have to be specified in a more detailed way.

Now, for deriving steady state solutions, we set (A1) to zero. In this case, $\boldsymbol{g}(\boldsymbol{t})$ has to be a constant g_0 and one gets

(A2) $$\boldsymbol{y_B}(\boldsymbol{t}) = \frac{\boldsymbol{g_0}}{\boldsymbol{c_y}}$$ and

(A3) $$\boldsymbol{y_U}(\boldsymbol{t}) = \frac{\boldsymbol{g_0}}{\boldsymbol{v}}.$$

by inserting (A2) into equation (1) in the main text. This indicates that for analytes being in *steady state* the log analyte concentration should be linear to the log of the urinary flow rates. Note that the steady state solutions derived here are commonly used to calculate fractional excretions [27], so we are moving on common ground. For metabolites not in steady states, we have to solve (A1) which is easily done by the standard technique of variation of the

parameters as (A1) is a linear differential equation. The linear differential equation (A1) thus is unambiguously solved by

$$\text{(A4)} \qquad \boldsymbol{y_B(t)} = \mathrm{e}^{-\mathrm{c_y}\boldsymbol{t}}\left[\mathrm{y_0} + \int_0^{\boldsymbol{t}} \mathrm{e}^{\mathrm{c_y}\boldsymbol{\xi}}\boldsymbol{g}(\boldsymbol{\xi})d\boldsymbol{\xi}\right].$$

Using the mean value theorem for integration with $\bar{\boldsymbol{g}}$ denoting the average net influx into blood in the interval [0, $\boldsymbol{t}$], (A4) simplifies to

$$\text{(A5)} \qquad \boldsymbol{y_B(t)} = \left(\mathrm{y_0} - \frac{\bar{\boldsymbol{g}}}{\mathrm{c_y}}\right)\mathrm{e}^{-\mathrm{c_y}\boldsymbol{t}} + \frac{\bar{\boldsymbol{g}}}{\mathrm{c_y}}.$$

Insertion of (A5) into equation (1) of the main text results in

$$\text{(A6)} \qquad \boldsymbol{y_U(t)} = \frac{\left(\mathrm{y_0} - \frac{\bar{\boldsymbol{g}}}{\mathrm{c_y}}\right)(1-\mathrm{e}^{-\mathrm{c_y}\boldsymbol{t}}) + \bar{\boldsymbol{g}}\mathrm{t}}{\mathrm{v}\boldsymbol{t}}.$$

Modeling the normalization process

In steady state, the urinary concentration can be seen as a direct indicator of the parameter $\bar{\boldsymbol{g}}$ if a good estimation $\hat{\mathrm{v}}$ of v is available. Typically, urine analyte data are normalized by dividing the raw concentration by an estimator $\hat{\mathrm{v}}$ of the urinary flow rate, inversely proportional to the true urinary flow rate with a constant k. Hence, division of (A6) by a normalization variable leads to

$$\text{(A7)} \qquad \boldsymbol{y_{u_N}(t)} = \left[\frac{\left(\mathrm{y_0} - \frac{\bar{\boldsymbol{g}}}{\mathrm{c_y}}\right)(1-e^{-\mathrm{c_y}\boldsymbol{t}})}{\boldsymbol{t}} + \bar{\boldsymbol{g}}\right]\frac{\hat{\mathrm{v}}}{\mathrm{v}} = \left[\frac{\left(\mathrm{y_0} - \frac{\bar{\boldsymbol{g}}}{\mathrm{c_y}}\right)(1-\mathrm{e}^{-\mathrm{c_y}\boldsymbol{t}})}{\boldsymbol{t}} + \bar{\boldsymbol{g}}\right]\frac{1}{\mathrm{k}}.$$

The constant k should be equal across individuals and over time, otherwise the results are systematically biased through normalization. Creatinine normalization (assuming creatinine concentration in blood is in steady state) is an example of a normalization procedure introducing bias as the constant k is then equal to the creatinine production rate, which is thought to be a proxy for muscle mass, varying with sex, age, and body mass index [20]. As

(A7) is a function of $\boldsymbol{t}$ for non-steady state analytes, the normalized urinary concentrations are a function of the length of the time interval between the urine voids. For example, for constant net influxes, (A7) will be decreasing with increasing time interval if the starting value would be above the steady state solution.

Transferring the model to the population level

In reality, we do not have continuous analyte measurements over time of one person, but only point measurements of many persons at different times. Therefore, we have to think of the parameters of equation (A7) as random variables. We rewrite (A7) in the following way

$$\text{(A8)} \qquad \boldsymbol{Y_{U_N}} = \left[\left(\boldsymbol{Y_0} - \frac{\bar{\boldsymbol{G}}}{\boldsymbol{C_y}}\right)(1 - \mathrm{e}^{-\boldsymbol{C_y T}})\boldsymbol{T^{-1}} + \bar{\boldsymbol{G}}\right]\boldsymbol{V^{-1}}\widehat{\boldsymbol{V}}\,,$$

where the capitals are representing the corresponding random variables to the before made definitions of the characters. A urinary concentration is thus a function of the random variables vector $(\boldsymbol{Y_0}, \boldsymbol{C_y}, \bar{\boldsymbol{G}}, \boldsymbol{T}, \boldsymbol{V}, \widehat{\boldsymbol{V}})$ which has an unknown distribution and a specific covariance matrix for a certain population. Each observation $Y_{u_{N_i}}$ is a realization of (A8) and a sample of urinary measurements can be seen as independent and identically distributed (iid) realizations of the vector $(\boldsymbol{Y_0}, \boldsymbol{C_y}, \bar{\boldsymbol{G}}, \boldsymbol{T}, \boldsymbol{V}, \widehat{\boldsymbol{V}})$**.** The urinary flow rate logically depends negatively on the time variable as explained above. This implies that the normalization variable covariates with the time interval variable which in return means that even after normalization the data will not be stochastically independent of the dilution variable. This is further complicated by the fact that $\bar{\boldsymbol{G}}$ is also a function of $\boldsymbol{T}$ for analytes with non-constant blood concentrations. It follows that the normalization variable will not be log linear to the raw urinary concentration if an analyte is not in steady state *population-wise*.

References

[1] G. Echeverry, G.L. Hortin, A.J. Rai, Introduction to urinalysis: historical perspectives and clinical application, Methods Mol. Biol. 641 (2010) 1–12. doi:10.1007/978-1-60761-711-2_1.

[2] S. Bouatra, F. Aziat, R. Mandal, A.C. Guo, M.R. Wilson, C. Knox, T.C. Bjorndahl, R. Krishnamurthy, F. Saleem, P. Liu, Z.T. Dame, J. Poelzer, J. Huynh, F.S. Yallou, N. Psychogios, E. Dong, R. Bogumil, C. Roehring, D.S. Wishart, The human urine metabolome, PLoS ONE. 8 (2013) e73076. doi:10.1371/journal.pone.0073076.

[3] I.F. Duarte, S.O. Diaz, A.M. Gil, NMR metabolomics of human blood and urine in disease research, J Pharm Biomed Anal. 93 (2014) 17–26. doi:10.1016/j.jpba.2013.09.025.

[4] S.M. Kohl, M.S. Klein, J. Hochrein, P.J. Oefner, R. Spang, W. Gronwald, State-of-the art data normalization methods improve NMR-based metabolomic analysis, Metabolomics. 8 (2012) 146–160. doi:10.1007/s11306-011-0350-z.

[5] A.-H. Emwas, C. Luchinat, P. Turano, L. Tenori, R. Roy, R.M. Salek, D. Ryan, J.S. Merzaban, R. Kaddurah-Daouk, A.C. Zeri, G.A. Nagana Gowda, D. Raftery, Y. Wang, L. Brennan, D.S. Wishart, Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review, Metabolomics. 11 (2015) 872–894. doi:10.1007/s11306-014-0746-7.

[6] Y. Gagnebin, D. Tonoli, P. Lescuyer, B. Ponte, S. de Seigneux, P.-Y. Martin, J. Schappler, J. Boccard, S. Rudaz, Metabolomic analysis of urine samples by UHPLC-QTOF-MS: Impact of normalization strategies, Anal. Chim. Acta. 955 (2017) 27–35. doi:10.1016/j.aca.2016.12.029.

[7] H.U. Zacharias, T. Rehberg, S. Mehrl, D. Richtmann, T. Wettig, P.J. Oefner, R. Spang, W. Gronwald, M. Altenbuchinger, Scale-Invariant Biomarker Discovery in Urine and Plasma Metabolite Fingerprints, Journal of Proteome Research. 16 (2017) 3596–3605. doi:10.1021/acs.jproteome.7b00325.

[8] B.M. Warrack, S. Hnatyshyn, K.-H. Ott, M.D. Reily, M. Sanders, H. Zhang, D.M. Drexler, Normalization strategies for metabonomic analysis of urine samples, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 877 (2009) 547–552. doi:10.1016/j.jchromb.2009.01.007.

[9] J. Hochrein, H.U. Zacharias, F. Taruttis, C. Samol, J.C. Engelmann, R. Spang, P.J. Oefner, W. Gronwald, Data Normalization of (1)H NMR Metabolite Fingerprinting Data Sets in the Presence of Unbalanced Metabolite Regulation, J. Proteome Res. 14 (2015) 3217–3228. doi: 10.1021/acs.jproteome.5b00192.

[10] E. Saccenti, Correlation Patterns in Experimental Data Are Affected by Normalization Procedures: Consequences for Data Analysis and Network Inference, J. Proteome Res. 16 (2017) 619–634. doi: 10.1021/acs.jproteome.6b00704.

[11] A. Zhang, H. Sun, X. Wu, X. Wang, Urine metabolomics, Clin. Chim. Acta. 414 (2012) 65–69. doi:10.1016/j.cca.2012.08.016.

[12] D.R.S. Middleton, M.J. Watts, R.M. Lark, C.J. Milne, D.A. Polya, Assessing urinary flow rate, creatinine, osmolality and other hydration adjustment methods for urinary biomonitoring using NHANES arsenic, iodine, lead and cadmium data, Environ Health. 15 (2016) 68. doi:10.1186/s12940-016-0152-x.

[13] J.-F. Sauvé, M. Lévesque, M. Huard, D. Drolet, J. Lavoué, R. Tardif, G. Truchon, Creatinine and specific gravity normalization in biological monitoring of occupational exposures, J Occup Environ Hyg. 12 (2015) 123–129. doi:10.1080/15459624.2014.955179.

[14] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics, Anal. Chem. 78 (2006) 4281–4290. doi:10.1021/ac051632c.

[15] E.J. Cone, Y.H. Caplan, F. Moser, T. Robert, M.K. Shelby, D.L. Black, Normalization of Urinary Drug Concentrations with Specific Gravity and Creatinine, Journal of Analytical Toxicology. 33 (2009) 1–7. doi:10.1093/jat/33.1.1.

[16] P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers?, J Chromatogr A. 1362 (2014) 194–205. doi:10.1016/j.chroma.2014.08.050.

[17] J. Hertel, S. Van der Auwera, N. Friedrich, K. Wittfeld, M. Pietzner, K. Budde, A. Teumer, T. Kocher, M. Nauck, H.J. Grabe, Two statistical criteria to choose the method for dilution correction in metabolomic urine measurements, Metabolomics. 13 (2017). doi:10.1007/s11306-017-1177-z.

[18] R.A. Kronmal, Spurious Correlation and the Fallacy of the Ratio Standard Revisited, Journal of the Royal Statistical Society. Series A (Statistics in Society). 156 (1993) 379. doi:10.2307/2983064.

[19] A.S. Levey, L.A. Stevens, C.H. Schmid, Y.L. Zhang, A.F. Castro, H.I. Feldman, J.W. Kusek, P. Eggers, F. Van Lente, T. Greene, J. Coresh, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration), A new equation to estimate glomerular filtration rate, Ann. Intern. Med. 150 (2009) 604–612.

[20] W.F. Clark, J.M. Sontrop, J.J. Macnab, R.S. Suri, L. Moist, M. Salvadori, A.X. Garg, Urine Volume and Change in Estimated GFR in a Community-Based Cohort Study, Clinical Journal of the American Society of Nephrology. 6 (2011) 2634–2641. doi:10.2215/CJN.01990211.

[21] C. Rose, A. Parker, B. Jefferson, E. Cartmell, The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology, Critical Reviews in Environmental Science and Technology. 45 (2015) 1827–1879. doi:10.1080/10643389.2014.1000761.

[22] T.D. Noakes, G. Wilson, D.A. Gray, M.I. Lambert, S.C. Dennis, Peak rates of diuresis in healthy humans during oral fluid overload, S. Afr. Med. J. 91 (2001) 852–857.

[23] S. Klahr, S.B. Miller, Acute Oliguria, New England Journal of Medicine. 338 (1998) 671–675. doi:10.1056/NEJM199803053381007.

[24] E.S. Lukacz, C. Sampselle, M. Gray, S. MacDiarmid, M. Rosenberg, P. Ellsworth, M.H. Palmer, A healthy bladder: a consensus statement: Consensus statement - a healthy bladder, International Journal of Clinical Practice. 65 (2011) 1026–1036. doi:10.1111/j.1742-1241.2011.02763.x.

[25] M.W. Taal, B.M. Brenner, F.C. Rector, eds., Brenner & Rector's the kidney, 9th ed, Elsevier/Saunders, Philadelphia, PA, 2012.

[26] F.A. Gotch, The current place of urea kinetic modelling with respect to different dialysis modalities, Nephrol. Dial. Transplant. 13 Suppl 6 (1998) 10–14.

[27] H. Schartum-Hansen, P.M. Ueland, E.R. Pedersen, K. Meyer, M. Ebbing, Ø. Bleie, G.F.T. Svingen, R. Seifert, B.E. Vikse, O. Nygård, Assessment of urinary betaine as a marker of diabetes mellitus in cardiovascular patients, PLoS ONE. 8 (2013) e69454. doi:10.1371/journal.pone.0069454.

[28] W.H. Greene, Econometric analysis, 5th ed, Prentice Hall, Upper Saddle River, N.J, 2003.

[29] F.E. Harrell, Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, Springer, New York, 2001.

[30] K.-Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika. 73 (1986) 13–22. doi:10.1093/biomet/73.1.13.

[31] B.S. Cade, B.R. Noon, A Gentle Introduction to Quantile Regression for Ecologists, Frontiers in Ecology and the Environment. 1 (2003) 412. doi:10.2307/3868138.

[32] M. Rotter, S. Brandmaier, C. Prehn, J. Adam, S. Rabstein, K. Gawrych, T. Brüning, T. Illig, H. Lickert, J. Adamski, R. Wang-Sattler, Stability of targeted metabolite profiles of urine samples under different storage conditions, Metabolomics. 13 (2017). doi:10.1007/s11306-016-1137-z.

[33] S. Krug, G. Kastenmüller, F. Stückler, M.J. Rist, T. Skurk, M. Sailer, J. Raffler, W. Römisch-Margl, J. Adamski, C. Prehn, T. Frank, K.-H. Engel, T. Hofmann, B. Luy, R. Zimmermann, F. Moritz, P. Schmitt-Kopplin, J. Krumsiek, W. Kremer, F. Huber, U. Oeh, F.J. Theis, W. Szymczak, H. Hauner, K. Suhre, H. Daniel, The dynamic range of the human metabolome revealed by challenges, FASEB J. 26 (2012) 2607–2619. doi:10.1096/fj.11-198093.

[34] C. Menni, G. Zhai, A. Macgregor, C. Prehn, W. Römisch-Margl, K. Suhre, J. Adamski, A. Cassidy, T. Illig, T.D. Spector, A.M. Valdes, Targeted metabolomics profiles are strongly correlated with nutritional patterns in women, Metabolomics. 9 (2013) 506–514. doi:10.1007/s11306-012-0469-6.

[35] E. Daskalaki, G. Blackburn, G. Kalna, T. Zhang, N. Anthony, D.G. Watson, A study of the effects of exercise on the urinary metabolome using normalisation to individual metabolic output, Metabolites. 5 (2015) 119–139. doi:10.3390/metabo5010119.

[36] J.R. Everett, Pharmacometabonomics in humans: a new tool for personalized medicine, Pharmacogenomics. 16 (2015) 737–754. doi:10.2217/pgs.15.20.

[37] M. Wallace, Y.Z.H.-Y. Hashim, M. Wingfield, M. Culliton, F. McAuliffe, M.J. Gibney, L. Brennan, Effects of menstrual cycle phase on metabolomic profiles in premenopausal women, Hum. Reprod. 25 (2010) 949–956. doi:10.1093/humrep/deq011.

[38] E.C.-P. Chua, G. Shui, I.T.-G. Lee, P. Lau, L.-C. Tan, S.-C. Yeo, B.D. Lam, S. Bulchand, S.A. Summers, K. Puvanendran, S.G. Rozen, M.R. Wenk, J.J. Gooley, Extensive diversity in circadian regulation of plasma lipids and evidence for different circadian metabolic phenotypes in humans, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) 14468–14473. doi:10.1073/pnas.1222647110.

[39] S. Araki, F. Sata, K. Murata, Adjustment for urinary flow rate: an improved approach to biological monitoring, International Archives of Occupational and Environmental Health. 62 (1990) 471–477. doi:10.1007/BF00379066.

[40] T. Sorahan, D. Pang, N. Esmen, S. Sadhra, Urinary concentrations of toxic substances: an assessment of alternative approaches to adjusting for specific gravity, J Occup Environ Hyg. 5 (2008) 721–723. doi:10.1080/15459620802399997.

[41] F. Sata, S. Araki, K. Yokoyama, K. Murata, Adjustment of creatinine-adjusted values in urine to urinary flow rate: a study of eleven heavy metals and organic substances, Int Arch Occup Environ Health. 68 (1995) 64–68.

[42] B.D. Rose, H.G. Rennke, Renal pathophysiology: the essentials, Williams & Wilkins, Baltimore, 1994.

Tables and Figures

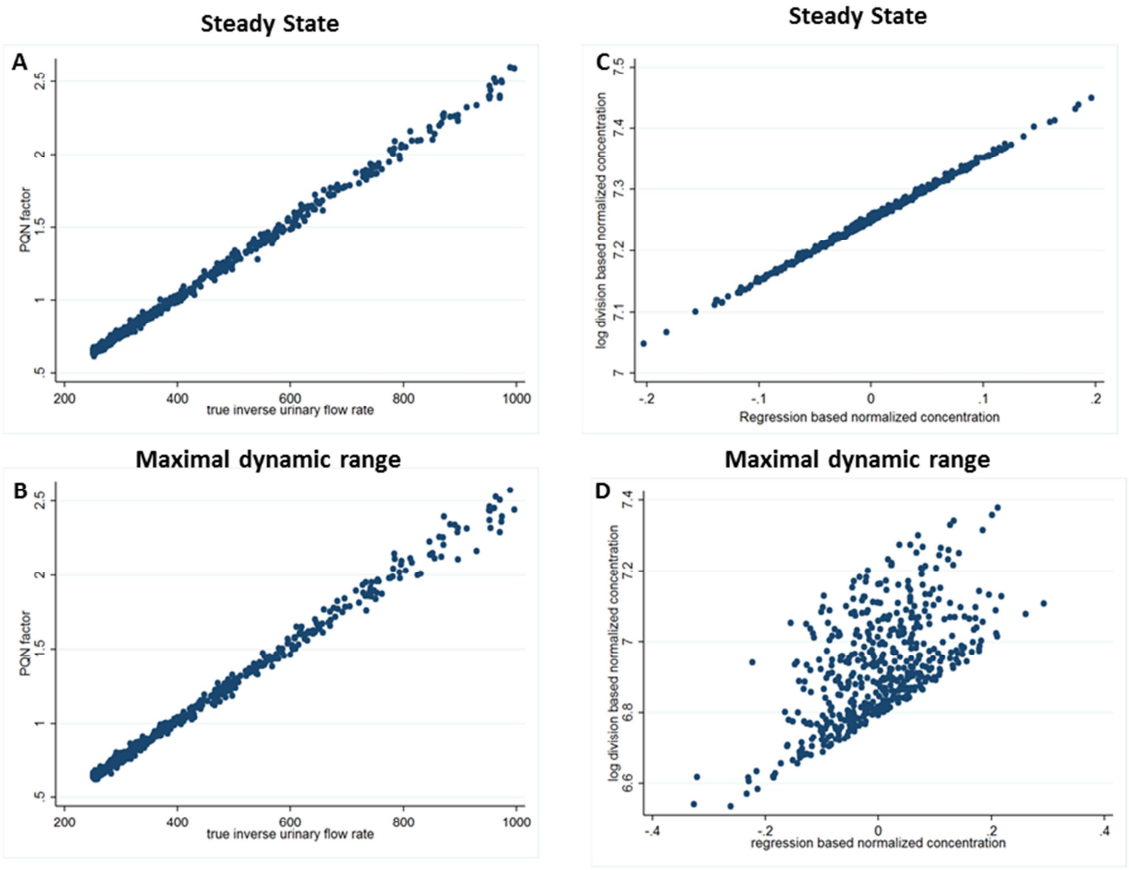


Figure 1: Results of the simulations. Figure 1A and 1B show the scatter plots of the PQN factor against the inverse urinary flow rate in steady state (A) and in the scenario of maximal dynamics (B). In both cases, the correlation was 0.99. Figure 1C and 1D show the exemplary scatter plot of regression-based normalizations against division-based normalizations. In the steady state case (C), the correlation was 0.99, while in the case of maximal dynamics (D) the correlation for this metabolite was only 0.60.

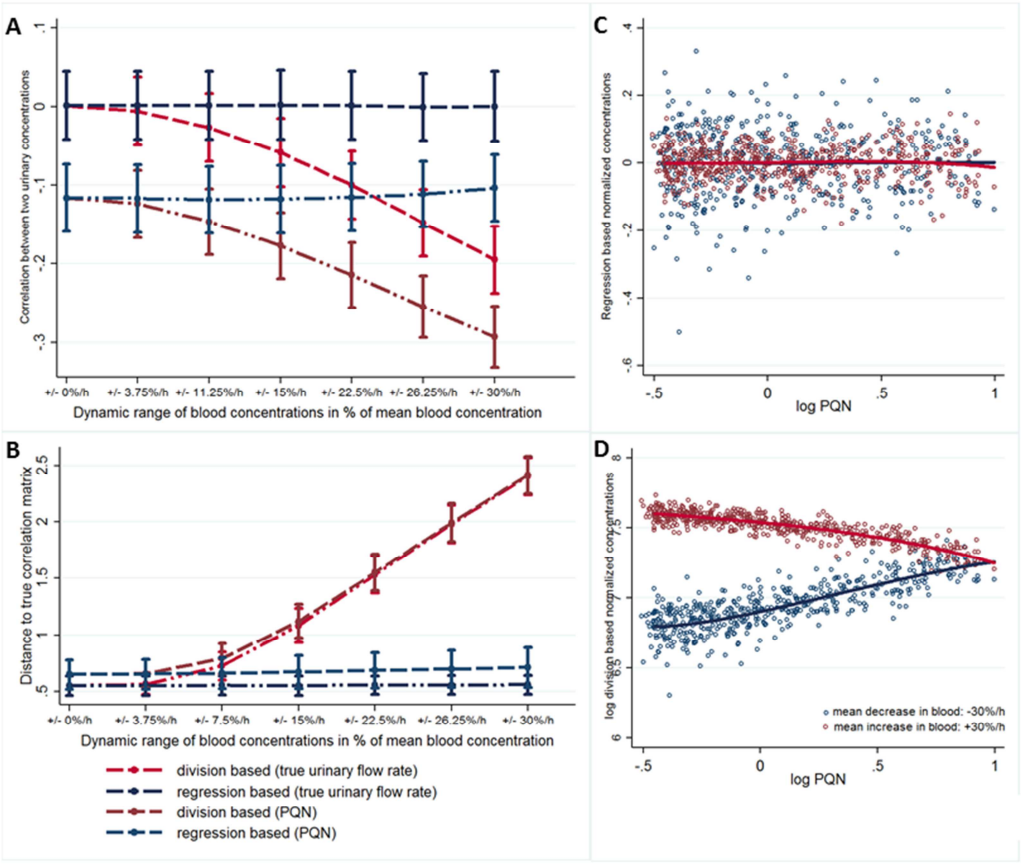


Figure 2: Results of the simulations. Error bars in Figure 2A and 2B display standard deviations from 500 independent replicates. In Figure 2A, the correlation of two theoretically independent metabolites is shown in dependency of the dynamic range in blood for division- and regression-based normalizations using the true urinary flow rate or the PQN factor as normalization factor, respectively. Figure 2B displays the distance (Frobenius norm) of the correlation matrix of the normalized data to the true correlation matrix of data devoid of variation in dilution. Figure 2C and figure 2D are normalization diagnostic graphs, showing the dependency of normalized concentrations on the normalization factor for regression-based normalization (C) and division-based normalization (D).

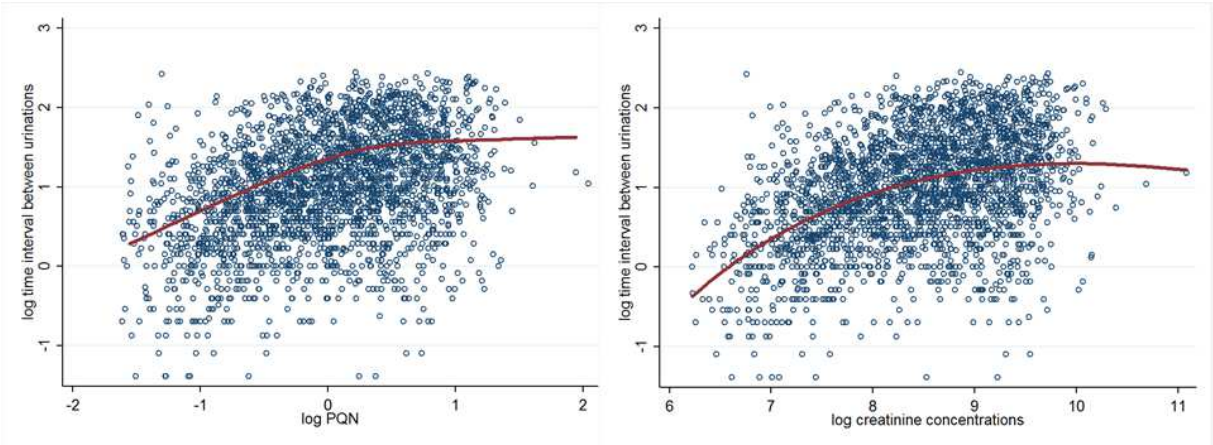


Figure 3: Scatter plots of the time interval between urinations against different normalization factors with flexible non-linear fits. The PQN variable explained 18.52% (95%-CI: (15.49%, 21.55%), p=7.844e-29 bootstrap derived using 2000 replicates of resampling) of intra-individual variance in the time interval between urinations and the creatinine concentrations explained 20.99% (95%-CI: (17.84%,24.13%), bootstrap derived using 2000 replicates of resampling).

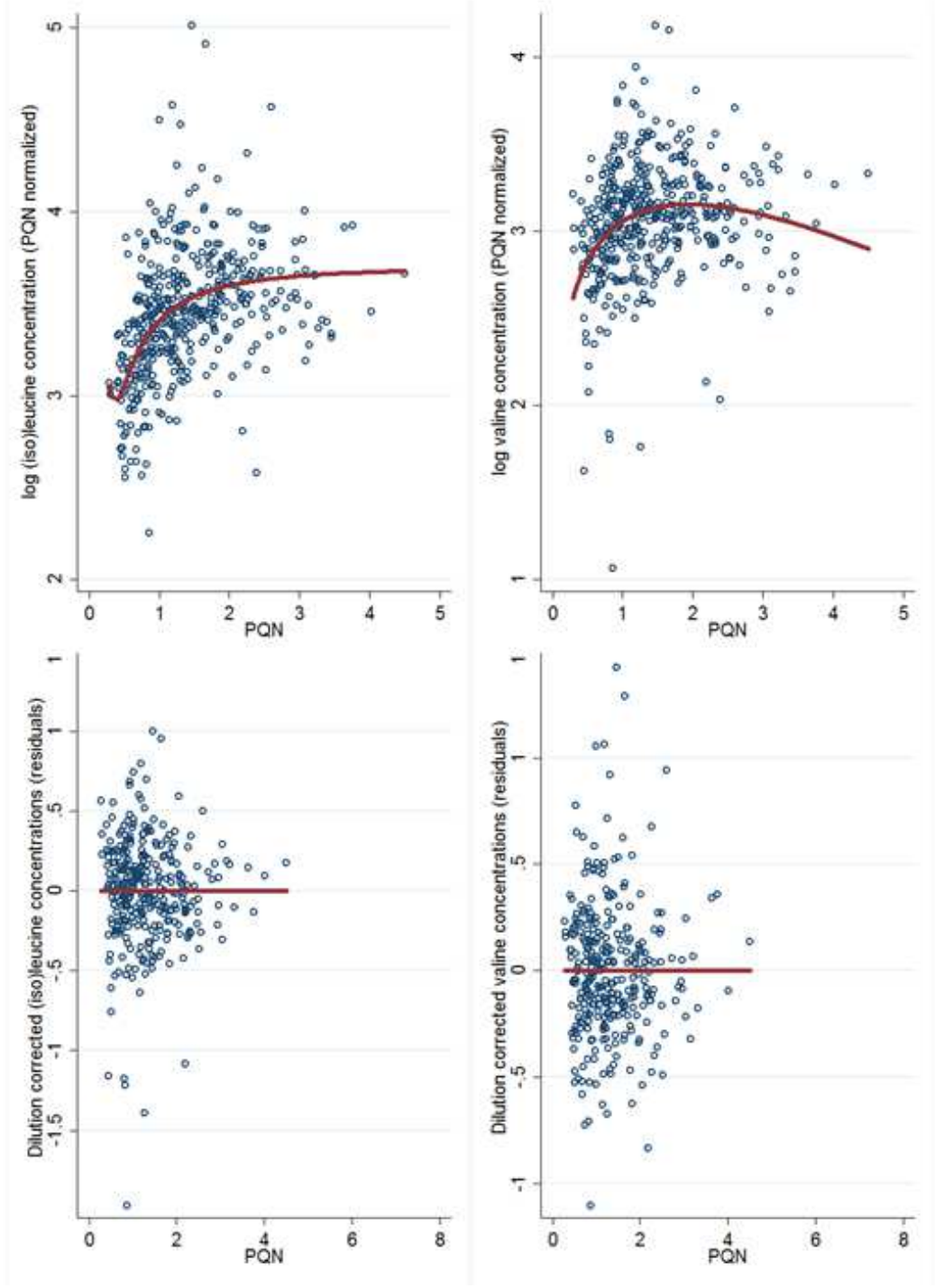


Figure 4: Examples of normalization diagnostics graphs. a) Non-linear dependencies of division-based normalized leucine/isoleucine and valine concentrations on the normalization factor for the wakeup urine sub-sample. b) Scatter-plot and regression line after non-linear correction (regression-based normalization) for the normalization factor.

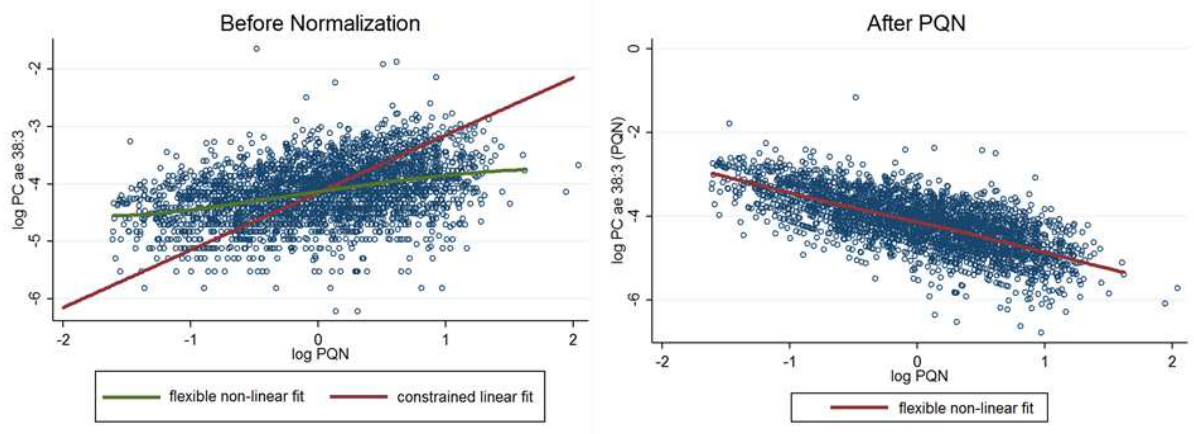


Figure 5: Dependency of PC ae C38:3 on the PQN factor before and after the division based PQN. The correlation to the log PQN factor is stronger after normalization (r=0.38 vs. r=-0.69)

Table 1: Summary of the results regarding the PQN and the predictions from the theoretical modeling

Metabolite	Dependence on Time-span since last urine void			Dependence on PQN before normalization	Dependence on PQN after normalization	
	p-value overall	p-value for departure from linearity	Shape of association	p-value for departure from log-linearity	p-value overall	Shape of association
Acylcarnitines						
C0	0.424	0.248	-	0.4351	0.0001	∩
C10	3.32e-05	0.106	↓	1.15e-20	2.3e-131	↓
C10:1	4.09e-08	1.026e-07	∩	0.0393	1.20e-31	↑
C10:2	0.341	0.633	-	0.0402	7.46e-08	↑
C12	0.0002	0.095	↓	9.23e-19	6.76e-64	↓
C14	3.74e-30	8.99e-06	↓	3.19e-95	3.4e-274	U
C14:1	0.302	0.517	-	1.72e-35	1.59e-41	U
C14:1 OH	2.47e-17	0.291	↓	1.34e-56	1.3e-128	U
C14:2	0.082	0.157	-	5.29e-60	2.73e-65	U
C14:2 OH	4.458e-34	1.154e-06	↓	2.82e-74	6.4e-263	U
C16	9.030e-10	0.821	↓	4.38e-33	7.53e-57	U
C16 OH	6.916e-07	0.405	↓	2.31e-32	1.16e-56	U
C16:2	7.026e-35	2.045e-10	↓	7.97e-62	1.00e-300	U
C18:2	1.642e-52	1.194e-11	↓	1.05e-52	1.00e-300	↓
C2	0.007	0.368	↑	0.0593	8.39e-24	↑
C3	0.247	0.172	-	0.0185	1.63e-11	↑
C4:1	0.003	0.141	↑	5.71e-06	3.95e-46	↑
C5	0.515	0.321	-	0.6387	0.6074	-
C5 MDC	0.018	0.413	↑	0.0001	9.38e-08	↑
C5:1	6.345e-06	0.979	↑	1.37e-05	4.28e-06	↓
C5:1 DC	1.892e-14	0.007	↑	1.43e-09	2.58e-80	∩
C6:1	0.0010	0.0008	U	0.0002	3.52e-48	↓
C7 DC	7.479e-11	2.039e-07	∩	1.12e-06	8.71e-39	↑
C8	0.004	0.014	U	3.55e-26	1.54e-29	↓
C8:1	1.996e-09	0.066	↑	0.0012	1.80e-67	↑
C9	2.889e-14	1.543e-12	∩	0.4908	4.15e-15	∩
Amino Acids						
ARG	2.138e-07	0.496	↑	3.97e-10	1.99e-50	∩
GLN	0.002	0.949	↑	1.27e-43	8.66e-43	∩
GLY	1.15e-05	0.294	↑	2.68e-19	3.32e-38	∩
HIS	7.896e-07	0.604	↑	1.56e-50	1.08e-52	∩
MET	0.018	0.944	↑	4.90e-30	6.35e-40	∩
PHE	0.0012	0.455	↓	1.17e-43	8.10e-56	∩
PRO	2.128e-38	4.957e-07	↑	8.56e-33	1.00e-300	↑
SER	0.0004	0.281	↑	5.93e-25	2.54e-43	∩
THR	2.322e-08	0.176	↑	6.83e-21	1.19e-67	∩
TRP	0.042	0.180	↑	3.38e-35	8.45e-35	∩
TYR	2.907e-05	0.008	∩	2.20e-26	6.9e-112	∩
VAL	6.257e-09	0.679	↑	1.65e-28	1.99e-82	∩
ILE/LEU	3.515e-23	0.017	↑	9.62e-26	4.4e-204	∩
CREATININE	1.299e-70	9.816e-12	↑	1.70e-27	1.00e-300	↑

Lipids						
PC ae C38:3	1.663e-30	2.889e-14	U	0.0467	1.00e-300	↓
PC ae C38:6	0.021	0.400	↑	0.0194	1.28e-08	↓
SMC24:0	0.0035	0.056	↓	1.62e-05	6.23e-52	↓
Hexoses						
H1	0.035	0.041	∩	0.0313	8.51e-33	∩

↓=monotonously decreasing; ↑ =monotonously increasing; ∩=inverse u-shaped association; U=u-shaped association; shape of association derived from graphical inspection; P-values from mixed models adjusting for age, bmi, batch, and the sampling time (non-linear via restricted cubic splines).

Table 2: Association of PC ae 38:3 to age (inter-individual factor) and time of day of urine sampling (intra-individual factor) using different normalization strategies

Association of PC ae 38:3 to age and time of day	b	SE of regression coefficient	ΔR-Squared	p-value
Age[1]			**Between**	
PQN (Division)	0.002	0.003	0.6%	0.472
PQN (RCS)	-0.003	0.002	1.21%	0.105
Integral (Division)	**0.008**	**0.004**	**5.02%**	**0.019**
Integral (RCS)	-0.002	0.002	0.47%	0.252
Creatinine (Division)	**0.106**	**0.003**	**8,41%**	**0.002**
Creatinine (RCS)	-0.002	0.002	0.18%	0.382
Time of Day[1,2]	-	-	**Within**	
PQN (Division)	-	-	**2.73%**	**5.429e-17**
PQN (RCS)	-	-	**11.42%**	**3.089e-75**
Integral (Division)	-	-	**1.40%**	**1.197e-08**
Integral (RCS)	-	-	**10.99%**	**4.745e-74**
Creatinine (Division)	-	-	**3.08%**	**3.493e-09**
Creatinine (RCS)	-	-	**10.78%**	**1.864e-72**

b=regression coefficient, SE=standard error, PC ae 38:3=phosphatidylcholine C38:3, PQN=probabilistic quotient normalizing, RCS=restricted cubic splines

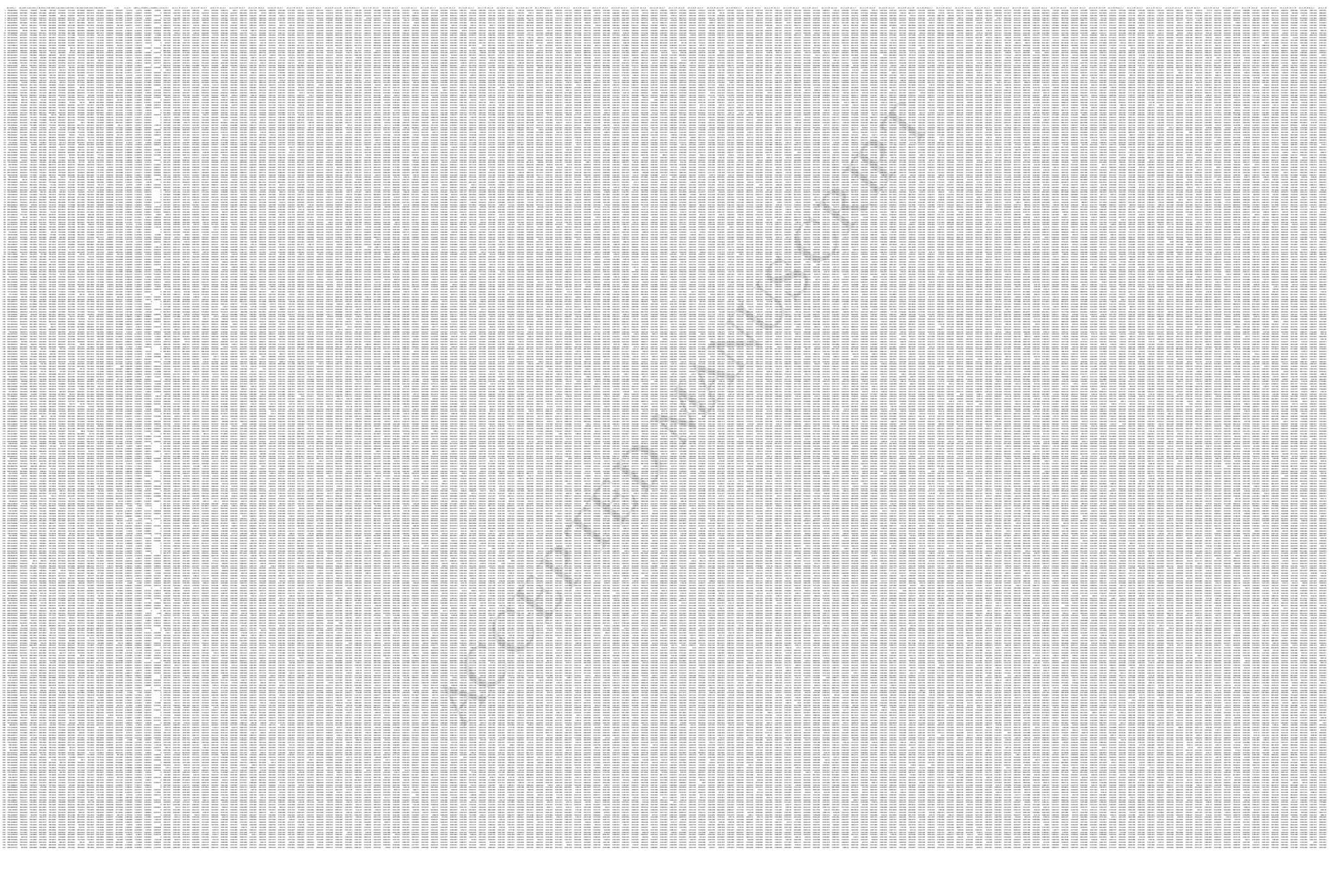
[1]Results from mixed linear regression, adjusted for Batch, random effects for individuals

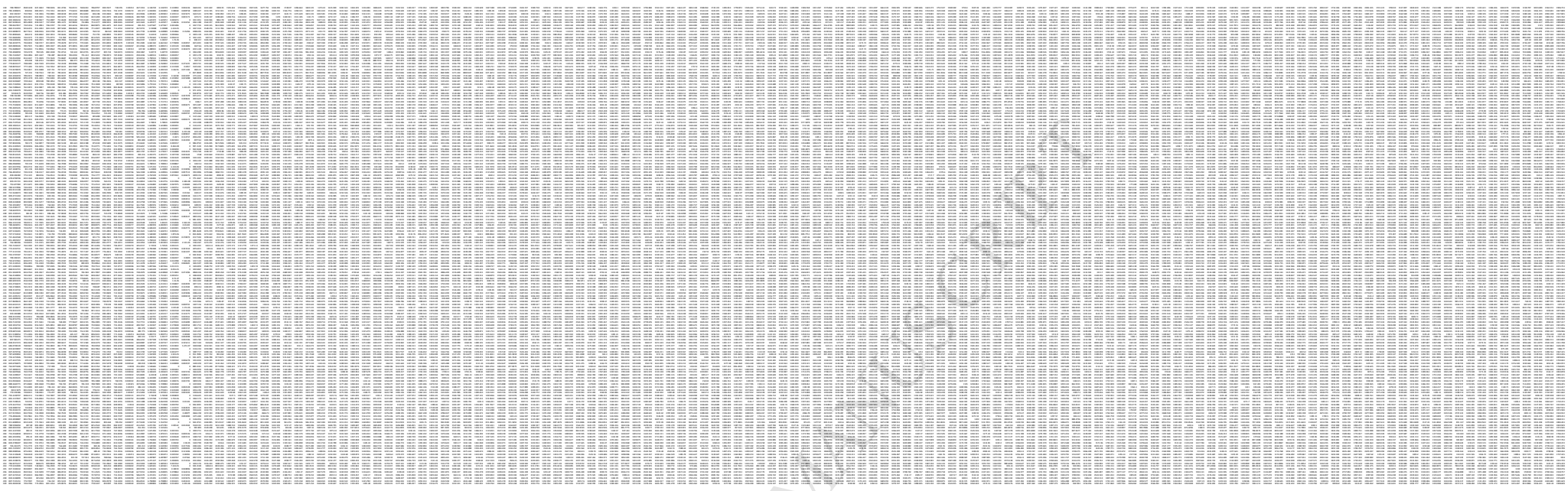
[2]p-value for association of time of day variable (restricted cubic splines using four knots) to PC ae 38:3 is derived by Wald test, testing all three spline variables simultaneously on zero. As the estimated functional relation of urinary metabolite concentration to time of day was inherently non-linear, regression coefficients are not interpretable and thus not shown.

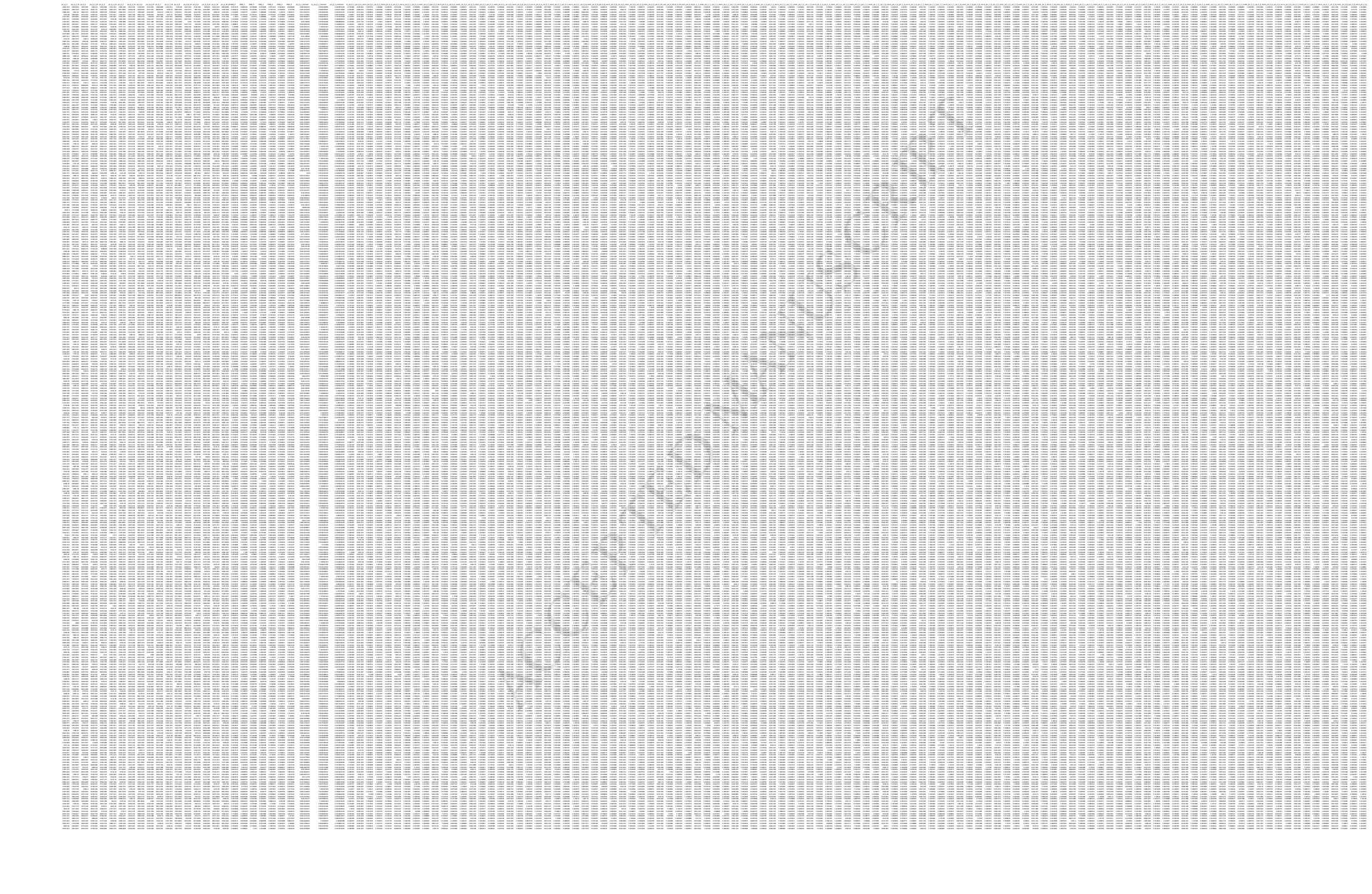
Data-dictionary

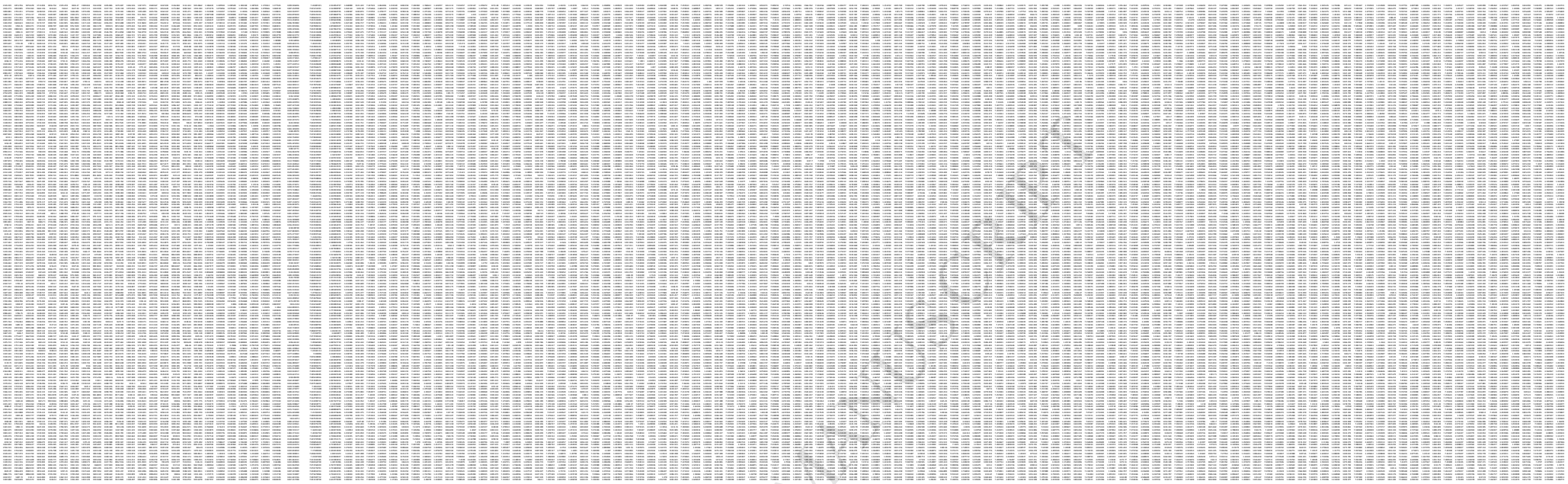
Variables	Label
id	Observation number
yB_mean_j	Mean blood concentrations in blood for the metabolites j=1,2,…,10
v	Urinary flow rate
ln_v_inv	log inverse urinary flow rate
spline_v_inv1-spline_v_inv3	spline transformation of the inverse urinary flow rate
yU_k_j	Urinary concentration for dynamic range k=0,1,…,6 and metabolite j=1,2,…,10
yU_k_j_t0	Urinary concentration for dynamic range k=0,1,…,6 and metabolite j=1,2,…,10 with constant urinary flow rate
PQN_k	PQN factor for dynamic range k
spline_PQN_k_1-spline_PQN_k_3	spline transformations of the log PQN for dynamic range k
yU_k_j_normed	Division-based normalized urinary concentrations (PQN) for dynamic range k=0,1,…,6 and metabolite j=1,2,…,10
yU_k_j_vnormed	Division-based normalized urinary concentrations (true urinary flow rate) for dynamic range k=0,1,…,6 and metabolite j=
yU_k_j_rcs	Regression-based normalized urinary concentrations (PQN) for dynamic range k=0,1,…,6 and metabolite j=1,2,…,10
yU_k_j_vrcs	Regression-based normalized urinary concentrations (true urinary flow rate) for dynamic range k=0,1,…,6 and metabolite

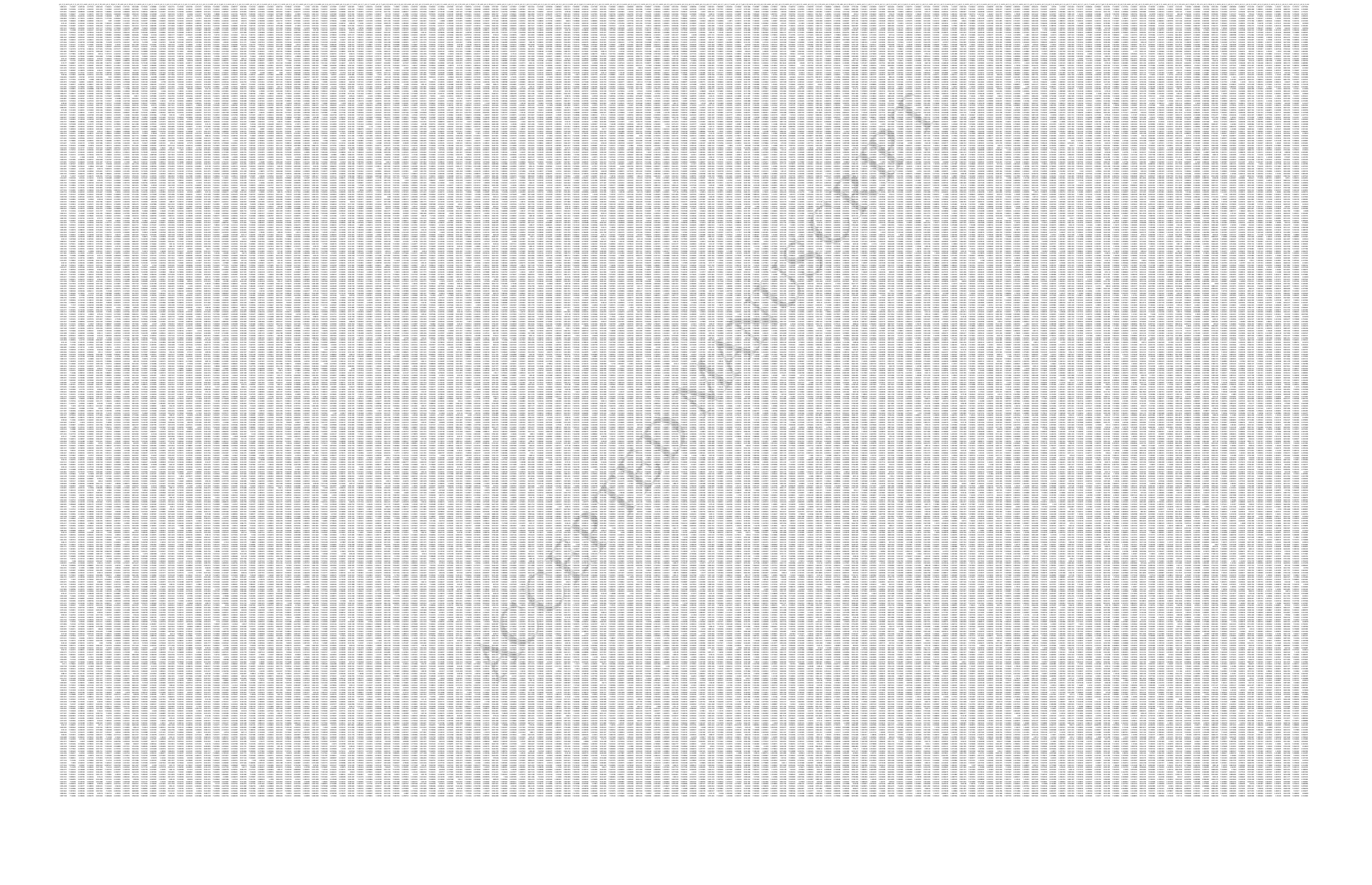
Parametrization of the variables is as described in the "Simulations" section in the supplementary material.

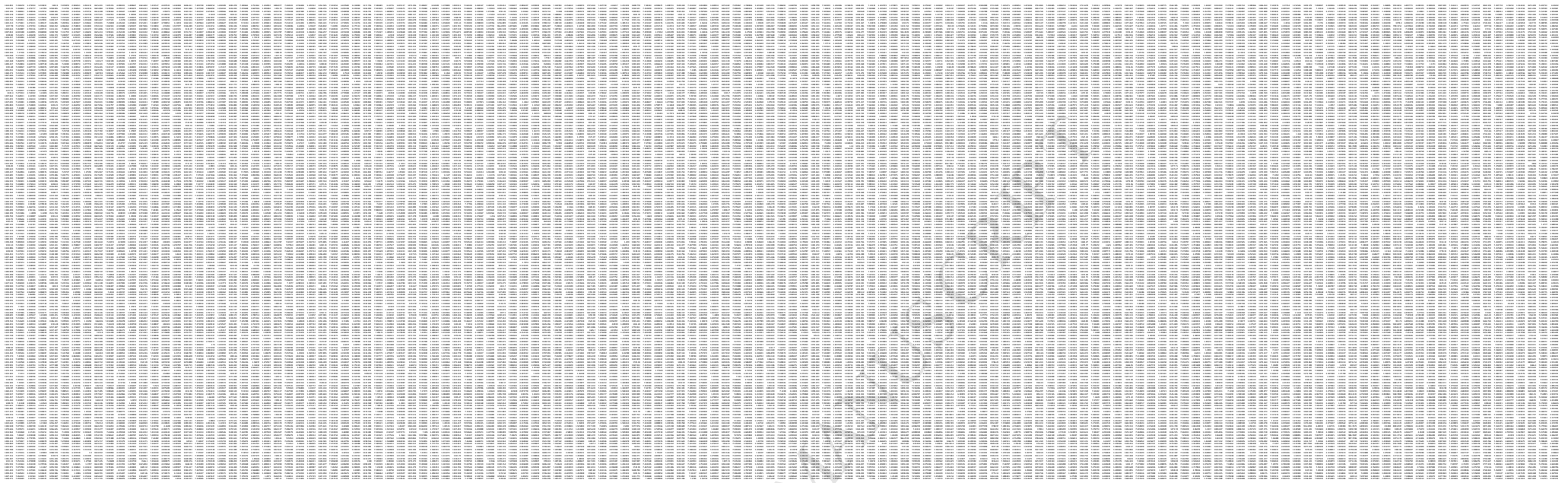


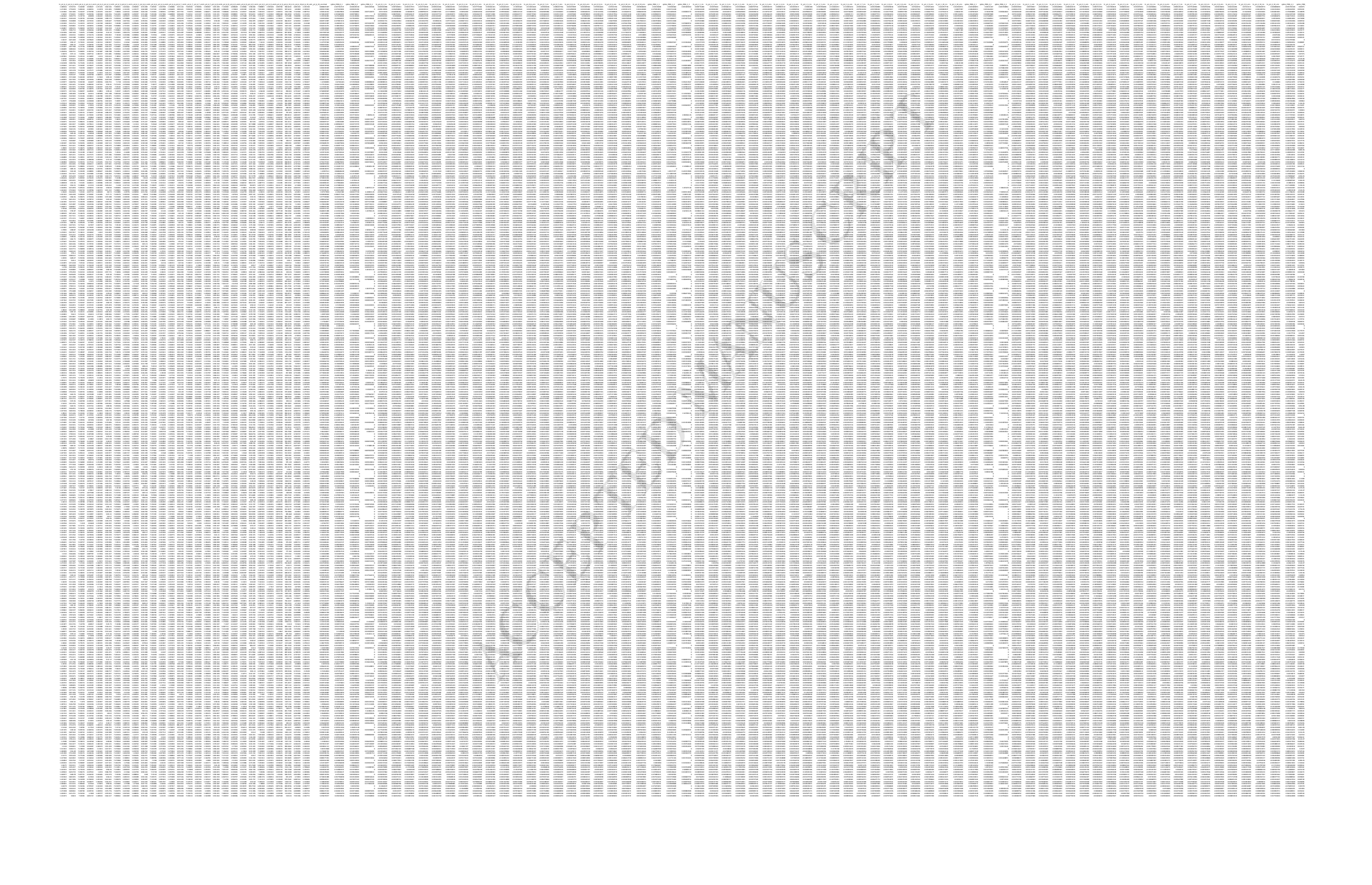


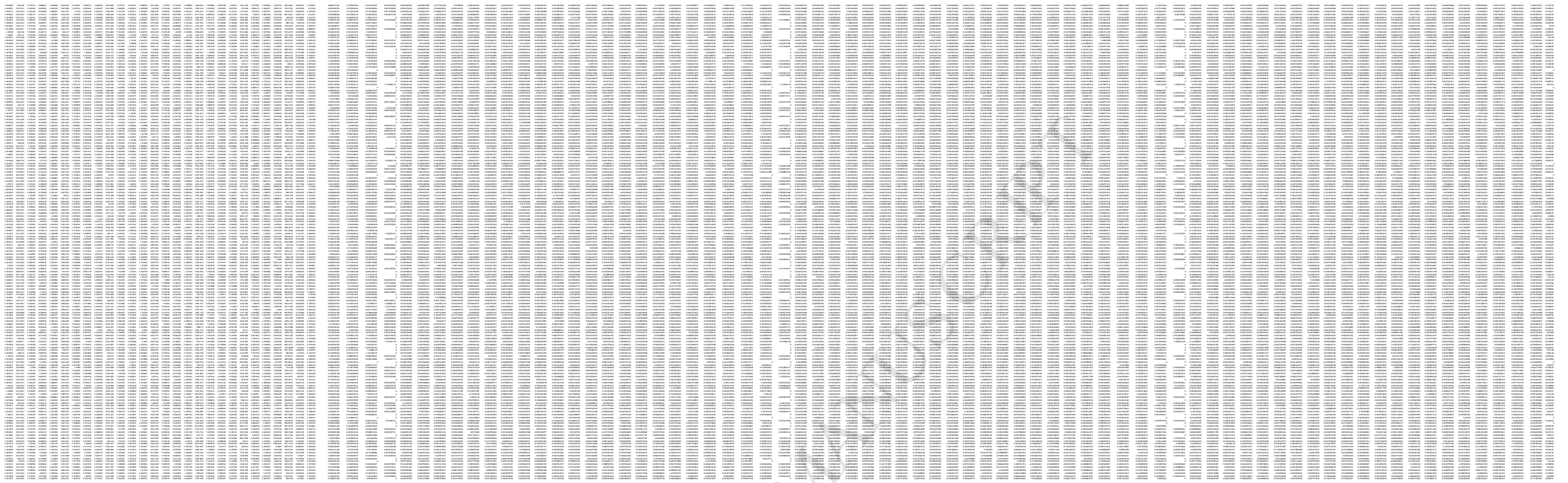


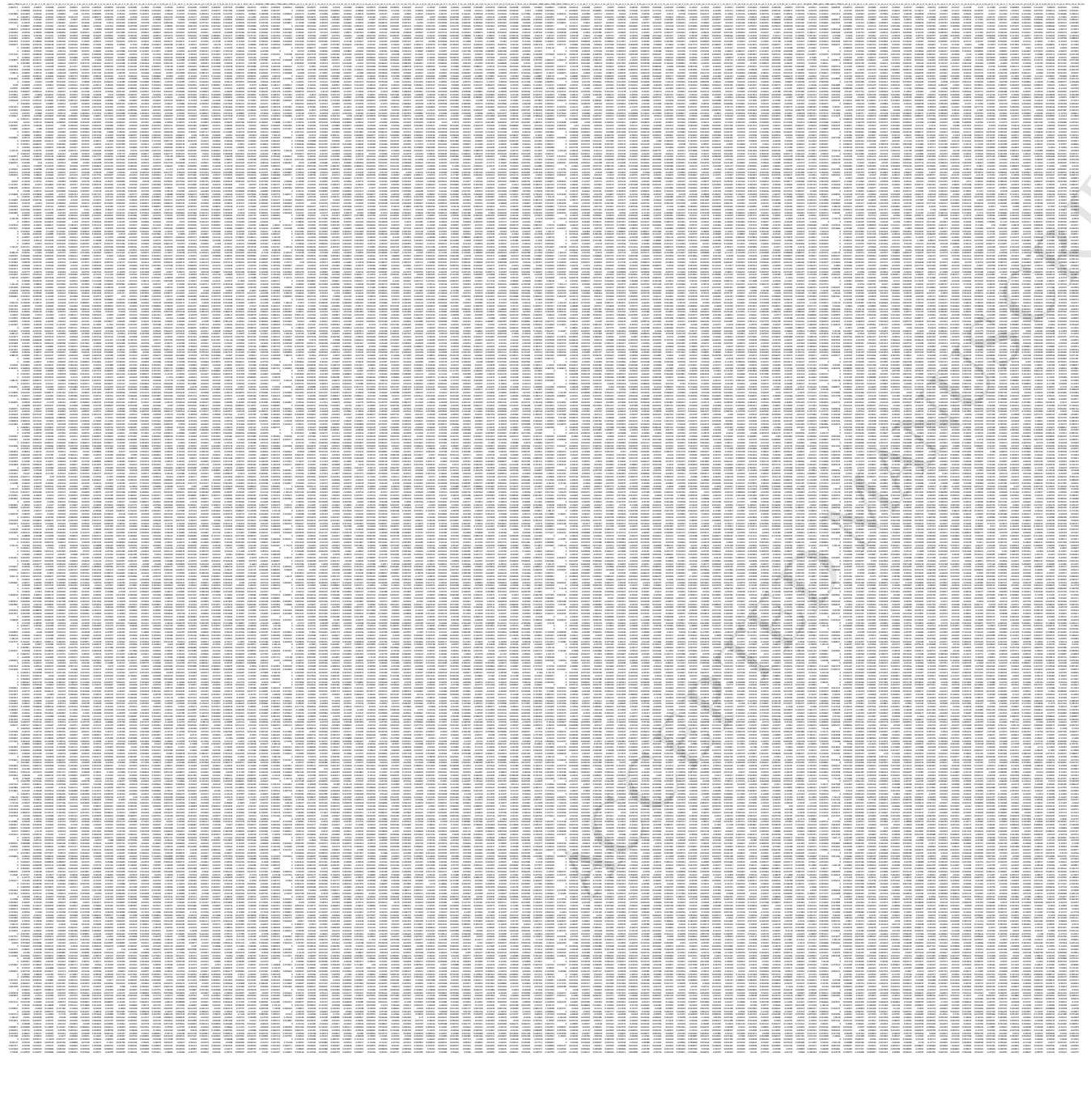


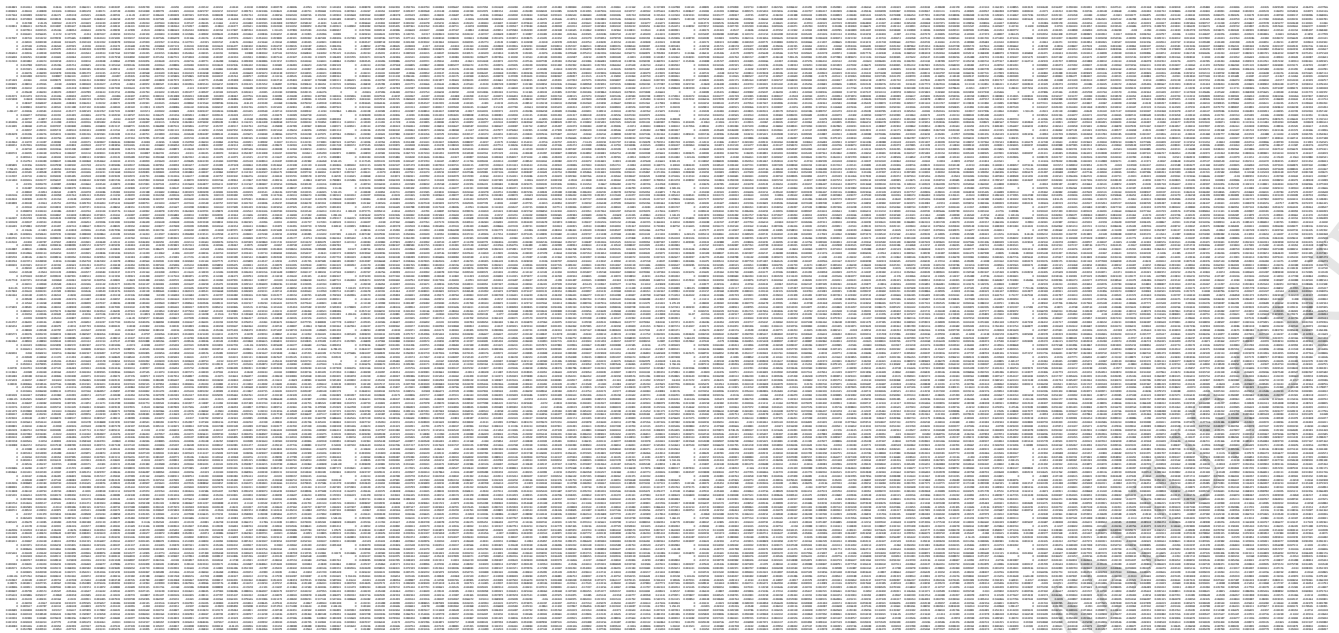


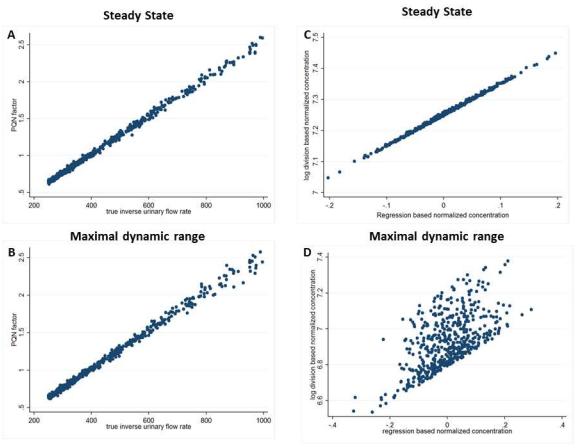


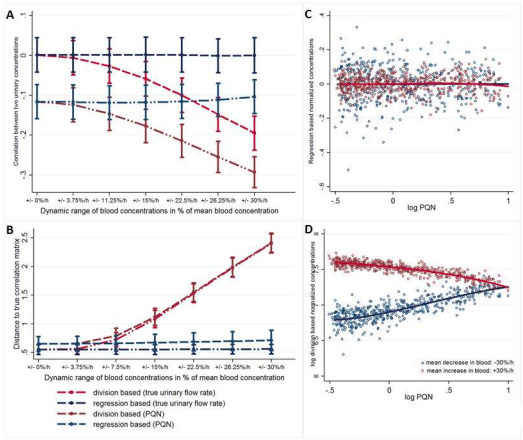


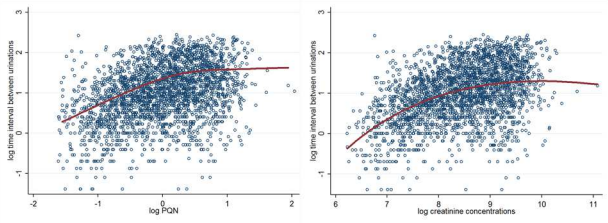


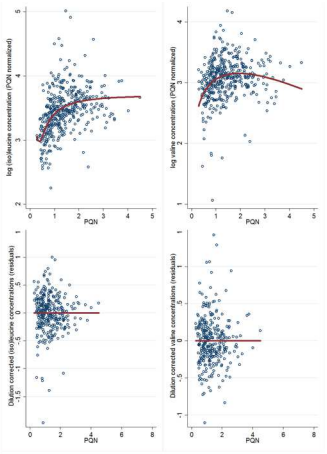


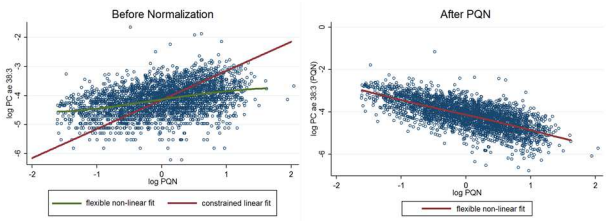












Highlights

- Division-based dilution correction implicitly assumes constant blood concentrations.
- The time-span since last urination is an important variance factor in urinary data.
- In dynamic data, each analyte needs its own dilution correction function.
- The efficiency of dilution correction can be tested via diagnostic graphs.
- Regression-based correction outperforms division-based methods in dynamic data.